Statistical Methods for the Spatio-Temporal Modelling of Infectious Disease Spread

---

Dissertation

zur
Erlangung der naturwissenschaftlichen Doktorwürde
(Dr. sc. nat.)

vorgelegt der
Mathematisch-naturwissenschaftlichen Fakultät
der
Universität Zürich

von

Sebastian Meyer

aus
Deutschland

Promotionskomitee

Prof. Dr. Leonhard Held (Vorsitz)
Prof. Dr. Reinhard Furrer
Prof. Dr. Michael Höhle
Prof. Dr. Torsten Hothorn
Prof. Dr. Mark Robinson

Zürich, 2016

Zusammenfassung

Infektionskrankheiten beeinträchtigen die öffentliche Gesundheit und erhalten besondere Aufmerksamkeit durch Epidemien. Saisonale Grippewellen und Norovirus-Ausbrüche zeigen, wie einfach sich mikroparasitäre Krankheitserreger ausbreiten können. Epidemische Modelle ermöglichen Vorhersagen und tragen zum Verständnis der Krankheitsausbreitung bei. Daher werden sie zunehmend für gesundheitspolitische Entscheidungen herangezogen.

Ursprünglich bestanden epidemische Modelle aus mathematischen Beschreibungen des Zeitverlaufs der Anzahl anfälliger, infizierter, und genesener Individuen einer Population. Durch etablierte Gesundheitsüberwachungssysteme werden jedoch räumlich aufgelöste Daten zum Auftreten von Infektionskrankheiten leichter verfügbar. Dies ermöglicht räumlich-zeitliche epidemische Modelle, statistische Inferenz, und letztendlich räumliche Vorhersagen. Durch Regressionsansätze kann beurteilt werden, wie Umweltfaktoren, sozioökonomische Merkmale oder Durchimpfungsraten die endemische und epidemische Verbreitung beeinflussen.

Aufbauend auf bestehenden Modellierungsansätzen für räumlich-zeitliche Überwachungsdaten erforscht diese Doktorarbeit, wie die Bevölkerungsstruktur und ihre Kontaktmuster besser berücksichtigt werden können. Insbesondere ist bekannt, dass menschliches Reiseverhalten einem Potenzgesetz in Bezug auf die zurückgelegte Distanz folgt. Zudem zeigen Studien zu sozialen Kontakten eine stark strukturierte Durchmischung der Altersgruppen.

Wir formulieren zunächst ein räumliches Potenzgesetz im Rahmen von zwei endemisch-epidemischen Modellen: ein räumlich-zeitlicher Punktprozess für Individualdaten und ein Zeitreihenmodell für regionale Fallzahlen. Anwendungen der so erweiterten Modelle auf Fälle invasiver Meningokokken-Erkrankungen bzw. Influenza-Fallzahlen zeigen, dass die Potenzgesetze tatsächlich durch die Daten gestützt werden und sich die Modellgüte und Vorhersagen deutlich verbessern.

In einem weiteren Schritt verallgemeinern wir das räumliche Zeitreihenmodell auf mehrfach geschichtete Zähldaten. Wir zeigen, wie eine empirisch abgeleitete Kontaktmatrix die Übertragungsgewichte im Modell steuern kann. In einer Fallstudie zur Ausbreitung von Norovirus-Gastroenteritis über Stadtteile und Altersgruppen führt die vorgeschlagene Kontaktstruktur zu einer Modellverbesserung gegenüber der sonst üblichen Annahme homogener Kontakte zwischen Altersgruppen.

Die Ätiologie einiger Krankheiten ist noch weitgehend unbekannt. Es ist unklar, ob soziale Kontakte oder räumliche Nähe zu einem Erkrankten das Risiko für die eigene Erkrankung erhöhen. In einer Fallstudie zu psychiatrischen stationären Aufnahmen entwickeln wir einen statistischen Test für Raum-Zeit-Interaktion, der in den obigen Punktprozess-Regressionsansatz eingebettet ist und somit an regionale sozioökonomische Merkmale angepasst werden kann.

Alle entwickelten statistischen Methoden sind in gut dokumentierten, quelloffenen R-Paketen implementiert. Die Doktorarbeit enthält auch einen anschaulichen Leitfaden, der die Anwendung der Methoden in anderen Projekten erleichtert.

Abstract

Infectious diseases affect public health and regularly gain attention during epidemics. Seasonal waves of influenza and norovirus outbreaks demonstrate the ease of epidemic spread of microparasitic pathogens. Epidemic models enable predictions and support the understanding of infectious disease spread. They are increasingly recognized as a useful tool to inform public health policies.

Epidemic models originally consisted of mathematical descriptions of the number of susceptible, infected, and recovered individuals in a population over time. However, spatial data on infectious disease occurrence are becoming more readily available from established public health surveillance systems. This enables spatio-temporal epidemic modelling, statistical inference, and thus spatial predictions. Regression approaches for such models allow us to assess the role of environmental factors, socio-economic characteristics, or vaccination coverage in shaping endemic and epidemic disease dynamics.

Building on existing modelling frameworks for spatio-temporal surveillance data, this thesis seeks methods to better account for population structure and contact patterns. In particular, it is known that human travel behaviour follows a power law with respect to distance. Furthermore, social contact surveys reveal a highly structured mixing of age groups.

We first incorporate a spatial power law in two endemic-epidemic models: a spatio-temporal point process for individual-level data and a time-series model for areal-level counts. Applying the extended models to cases of invasive meningococcal disease and counts of influenza, we show that the power-law formulations are indeed supported by the data and lead to substantial improvements of model fits and predictions.

In a further step, we adapt the areal time-series model to stratified count data. We show how an empirically-derived social contact matrix can inform transmission weights in the model. In a case study on the spread of norovirus gastroenteritis across city districts and age groups, the proposed contact structure outperforms the otherwise commonly used assumption of homogeneous mixing between age groups.

The etiology of some diseases is not yet well-established. It is unclear whether social contact or spatial proximity to a case elevates the risk of attracting the disease. In a case study on psychiatric inpatient admissions, we develop a statistical test for space-time interaction, which is embedded in the above point process regression approach and can thus be adjusted for regional socio-economic characteristics.

All developed statistical methods are implemented in well-documented open-source R packages. The thesis also contains an illustrative guide to facilitate application of the methods in other projects.

Thesis outline

Preface

This thesis is submitted under the Ph.D. program in "Epidemiology and Biostatistcis" at the University of Zurich. The thesis forms part of the research project "Statistical methods for spatio-temporal modelling and prediction of infectious diseases" funded by the Swiss National Science Foundation (project #137919), 2012–2015. The core output of this Ph.D. project are the four articles bundled in this thesis and accompanied by open-source implementations of the developed methods in readily usable R packages. An introductory chapter reviews the research topic and puts the articles in relation to each other.

Since scientific progress is not the result of a single person's work, I would like to thank all those who have supported my research, in particular ...

... Leonhard Held, who initiated this Ph.D. project, set the main research directions, and supervised my thesis over the whole time with regular and fruitful meetings. I am also grateful for the opportunity to participate in a summer school and to present my work at a dozen workshops and conferences. I appreciate that I was free to finish work from my previous research position in Munich and to dedicate parts of this thesis to the development of well-documented R packages implementing the proposed statistical methods.

... Michael Höhle, who can be clearly identified as the (infectious) source of my long-standing research in the statistical modelling of infectious disease surveillance data (he supervised both my Bachelor's and Master's Thesis at the Ludwig-Maximilians-Universität in Munich). We continued our collaboration in developing and promoting our favourite R package `surveillance`, and I am glad about my eventual inauguration as its "official" maintainer.

... the whole thesis committee for providing advice on this work.

... my colleagues at the Epidemiology, Biostatistics and Prevention Institute, especially my office mates Rafael Sauter, Daniel Sabanés Bové, and Rachel Heyard, for a nice working atmosphere.

... *An*drea Riebler and *Dan*iel Sabanés Bové for having established a Linux server and other useful IT services for our department, and Torsten Hothorn, for taking up the reins to rebuild our institute's IT infrastructure.

... and most prominently, my wife Julia and my two boys Jonathan and Moritz for offering hands-on experience with infectious disease spread, and for making sacrifices and cheering me up during work-intensive periods.

Zürich, July 2016

Introduction

Outbreaks of infectious diseases regularly alert public health. Authorities, physicians and epidemiologists are faced with questions on vaccination and its effectiveness, whether particular groups of the population are affected more than others, and how fast the disease is spreading. Emerging and re-emerging pathogens similarly threaten veterinary public health and plant ecology.

A useful tool to inform public health decisions during outbreaks are epidemic models (O'Neill, 2010). Examples include the investigations of Ferguson et al. (2001) during the 2001 UK foot-and-mouth epidemic and the work of the WHO Ebola Response Team (2014). These studies concentrate on providing rapid feedback on control measures and spatial projections for a specific outbreak. The epidemic models developed and applied in this thesis, however, are primarily intended for the retrospective analysis of long-term *spatio-temporal surveillance data*. These potentially contain multiple outbreaks, spatially heterogeneous incidence levels, and seasonal patterns with endemic periods. Thus, apart from being spatio-temporal, the other important feature of the models in this thesis is their *statistical* foundation. First of all, model parameters are being estimated from the data and uncertainty in these estimates and model predictions is quantifiable. This is in contrast to the vast literature devoted to deterministic (Anderson and May, 1991; Keeling and Rohani, 2008) or, more recently, computational (Balcan et al., 2009) models of infectious disease spread. Furthermore, by pursuing a regression approach, the proposed statistical models can relate data from external processes to disease spread. This is especially useful to quantify the role of environmental factors, seasonality, socio-demographic characteristics, vaccination, or other control measures in shaping endemic and epidemic dynamics.

This thesis pursues three aims: First, to establish statistical modelling frameworks for spatio-temporal surveillance data of epidemic phenomena in general. This is achieved by a thorough implementation in the `surveillance` package (Höhle et al., 2016) in R (R Core Team, 2016), accompanied by an illustrative guide (Paper II) to make the developed methods more accessible to others. Despite the intended widespread applicability of the models, the focus of this thesis is on human infectious diseases caused by directly transmitted microparasitic pathogens such as influenza or measles viruses. The corresponding second aim is to improve spatio-temporal epidemic models by incorporating empirically derived movement (Paper I) and social contact patterns (Paper III). Last but not least, Paper IV proposes a model-based test for "epidemicity", targeting diseases with unknown etiology.

The epidemic models in this thesis are thus intended to increase the understanding of infectious disease spread. That said, such statistical models are a prerequisite for probabilistic forecasts (cf. Paper I) and provide a basis for prospective outbreak detection (Diggle et al., 2005; Andersson et al., 2008; Manitz and Höhle, 2013; Piroutek et al., 2014; Salmon et al., 2016; Vial, Wei and Held, 2015). Prospective surveillance techniques are not further discussed in this work; suitable overviews can be found in the book edited by Lawson and Kleinman (2005) and in the review papers of Sonesson and Bock (2003) and Unkel et al. (2012).

The remainder of this introduction unfolds as follows: Section 1 reviews different approaches to epidemic modelling providing references to the corresponding literature. Afterwards, Section 2 gives a brief overview of the attached papers, which form the core output of this Ph.D. project.

1 Epidemic modelling

Several features of infectious disease surveillance data hinder the application of classical statistical approaches. For example, the data are rarely a result of planned experiments, the observations are not independent, and the epidemic process is only partially observable. The time point and location where and how an infection happened is usually unknown. Infections can also be asymptomatic or mild for some people, which then do not turn to a doctor and are not reported to public health authorities. A recent special issue in *Epidemics* (Lloyd-Smith et al., 2015) reviews these and other challenges in modelling infectious disease dynamics.

How epidemic modelling is approached depends on the purpose of the analysis. The forthcoming book chapter of Höhle (2016) is highly recommended for an introduction to various types of epidemic models and their relation – from deterministic, continuous-time, compartmental models for a single population to stochastic, discrete-time, spatial metapopulation models. The book chapter of Held and Paul

(2013) focusses on time series models for infectious disease counts, and Deardon et al. (2015) discuss individual-level approaches to epidemic modelling. In what follows, a brief review of the corresponding literature is provided.

1.1 Historical approaches

Traditionally, epidemic models have been deterministic, did not involve statistical inference to estimate parameters from actually observed data, and concentrated on the mathematical description of disease spread. Such deterministic models build on the work of Kermack and McKendrick (1927), who proposed a system of ordinary differential equations to describe the evolution of the number of *s*usceptible, *i*nfectious and *r*ecovered individuals in a fixed population over time, the so-called *SIR model*. The textbooks of Anderson and May (1991) and Keeling and Rohani (2008) cover this compartmental model and its extensions in great detail. Stochastic versions of the SIR model have been developed to account for the possibility of early extinction especially in small populations. Stochastic epidemic models are described in the textbooks of Becker (1989) [revisited by Becker and Britton (1999)], Daley and Gani (1999), and Andersson and Britton (2000). As recently surveyed by Britton (2010), many questions have been answered by studying mathematical properties of such models, including the threshold phenomenon of the *basic reproduction number*: if this average number of secondary cases directly caused by a single infective in a completely susceptible population is below one, an epidemic cannot take off. This concept has been used to guide public health policies, for example, by calculating the final size of an outbreak or by deriving the proportion of the susceptible population to vaccinate in order to prevent future outbreaks.

1.2 Spatial metapopulation models

However, considering the temporal dimension only and assuming a homogeneously mixing population of identical subjects does not allow for spatial predictions, and ignores the complex transmission dynamics in real populations, whether plant, animal or human. For plant diseases, the strength of interaction typically decays with distance, mainly as a result of aerodynamics and gravity for airborne diseases or as a result of vector dispersal (Madden et al., 2007; Cunniffe et al., 2015). The spread of livestock diseases is highly influenced by the movements of animals between farms (Fèvre et al., 2006; Brooks-Pollock et al., 2015). Human disease dynamics are largely driven by travel behaviour (Riley, 2007). To account for such distinctive structures, *metapopulation models* extend the basic SIR model for a single population to describe the epidemic spread within and between multiple sub-populations (Keeling et al., 2004). The most common type of

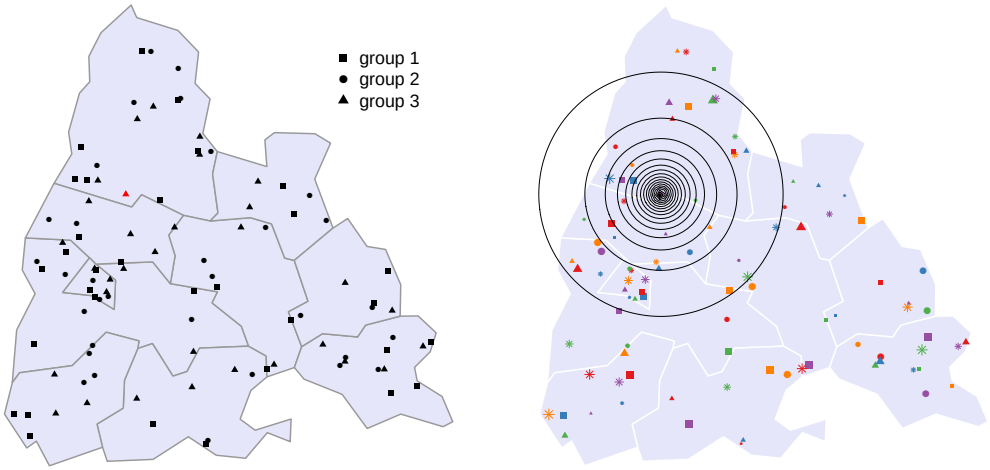
metapopulation model is the so-called *patch model*, where each subgroup refers to a spatially separated population, e.g., an administrative district or a farm.

Statistical approaches to modelling infectious disease spread across such patches have increasingly gained attention. Part of the reason is probably that the required spatial data has become more easily available from established infectious disease surveillance systems (Giesecke, 2002, Chapter 13). Human disease notification data typically form a time series of counts of reported cases, potentially stratified by region or age group. The Robert Koch Institute (RKI) in Germany, for example, makes such data routinely available as part of their *SurvStat@RKI* web service (`https://survstat.rki.de`) introduced in 2004 (Faensen and Krause, 2004).

Early work on time-series SIR (TSIR) models for count data is due to Finkenstädt and Grenfell (2000). It has since been extended to an *endemic*-epidemic model by Finkenstädt et al. (2002) to account for influx of new infections from unobserved sources, and to a spatially explicit version by Xia et al. (2004). However, these SIR-type models require the number of susceptibles to be known, which is unrealistic in regular surveillance settings. This inspired Held et al. (2005) to propose an endemic-epidemic time-series model only based on counts of new infections.

The crucial feature of metapopulation models is the strength of interaction between the different sub-populations: the "epidemiological coupling" (Keeling and Rohani, 2008, Chapter 7). In further applications of the model class of Held et al. (2005), coupling between regions has been measured by, for example, commuting flows and airline traffic for human influenza (Paul et al., 2008; Geilhufe et al., 2014), or cattle movement for livestock diseases (Schrödle et al., 2012). The spatial spread of influenza has been similarly studied by Viboud et al. (2006) and Eggo et al. (2011). However, the recent review of Chretien et al. (2014) reveals that only a few spatio-temporal approaches to influenza forecasting exist and that only Paul and Held (2011) and Held and Paul (2012) use proper scoring rules (Gneiting and Katzfuss, 2014) to validate their predictions. The influenza application in Paper I follows up on these works by incorporating parametric power-law weights to reflect human travel behaviour (Brockmann et al., 2006).

Various other statistical approaches have recently evolved, which all aim at modelling spatio-temporal interaction. For instance, Ensoy et al. (2013) investigate the spread of bluetongue among cattle farms, Aldrin et al. (2013) model the spread of salmon lice among salmon farms, Gog et al. (2014) investigate the 2009 H1N1 influenza pandemic in the USA, and Bauer et al. (2016) use a Bayesian spatio-temporal spline model to analyse hand-foot-and-mouth disease. In summarising current challenges for spatial epidemic models, Riley et al. (2015) accordingly state that "the explicit representation of space will likely become the norm rather than the exception for applied disease dynamics" – a development strongly supported by the readily usable model implementations illustrated in Paper II.



(a) Patch model with additional population strata. The red symbol refers to an infectious individual. The locations within each patch are irrelevant to this model.

(b) Distance-based transmission kernel, which may depend on individual characteristics.

Figure 1: Illustration of commonly assumed transmission rules in (a) metapopulation and (b) individual-level models. Adapted from Riley (2007, Figure 2).

1.3 Stratified patch models using social contact data

According to Ball et al. (2015), the metapopulation approach "has traditionally provided an attractive approach to incorporating more realistic contact structure into epidemic models". This is because the subgroups in metapopulation models may not only refer to spatial aggregations, but also to other population strata such as age groups, or to multiple levels of stratification (Cauchemez et al., 2011). Figure 1a shows the potential state of an epidemic as viewed by a stratified patch model. It is important to note that the exact positions of the individuals within a region are irrelevant for the assumed transmission rules in a patch model.

Further stratification is particularly relevant since the social phenomenon of "like seeks like" results in contact patterns between subgroups of a population, which go beyond a pure distance decay of interaction. Especially for school children, social contacts are known to be highly assortative with respect to age (Mossong et al., 2008). Characteristic pathways of directly transmitted pathogens are therefore likely to involve contacts between school children of the same age and their respective household members (Cauchemez et al., 2011). Eames et al. (2015) review challenges in measuring such contact networks, which shall reflect potential exposure to infectious agents in epidemic models. Social contact data have already been successfully used to estimate the age-dependent force of infection and

the basic reproduction number from seroprevalence data (Goeyvaerts et al., 2010; Hens et al., 2012). However, Paper III seems to be the first to employ social contact data in a spatio-temporal epidemic model. This works by extending the patch model of Paper I to an additional level of stratification, e.g., areal time series of counts further stratified by age group. As usual in metapopulation models, all susceptibles of a sub-population, such as group 1 in region 1, are equally likely to become infected. More specifically, the associated force of infection is determined by the contact structure between the groups, their relative distance in the neighbourhood graph of the regions, the prevalence of infection in each sub-population, and potentially further group- or region-specific characteristics.

1.4 Individual-level models

In individual-level models, the disease state and transmission rate are explicitly described for each individual of the population. The force of infection may thus depend on individual characteristics of the susceptible and infectious individuals and on the distance between them. Figure 1b illustrates a transmission kernel, where the strength of interaction between individuals decays with distance.

Individual-level statistical models of human diseases are often hindered by privacy protection issues, since case reports of human diseases are not commonly available, or their spatial resolution is limited. For example, the locations of the cases of invasive meningococcal disease used in Paper I and Paper II are only known up to the postcode level. In contrast, the patient data used in Paper IV is geo-referenced up to the building-level, and such high precision is standard for farm-level data on livestock diseases (Keeling et al., 2001) and for other epidemic phenomena. Many spatio-temporal epidemic models indeed go back to the Epidemic-Type Aftershock-Sequences (ETAS) point process model for earthquake occurrence (Ogata, 1998). This includes models for the spread of human infectious diseases (Meyer et al., 2012), invasive plant species in ecology (Balderama et al., 2012), and even for "predictive policing" (Mohler et al., 2015). These models describe the conditional intensity for an infection (event), given all past infections, by a superposition of endemic (background) and epidemic (triggering) components. The latter makes the process "self-exciting" in that each infectious individual (earthquake) triggers secondary infections (aftershocks, offspring) according to an inhomogeneous Poisson process – and these offspring events may induce further offspring, and so forth. The triggering rate typically decays with distance as well as over time, and may depend on individual-level characteristics (earthquake magnitude). For instance, Paper I adapts the endemic-epidemic model of Meyer et al. (2012) to use a power-law transmission kernel reflecting human travel behaviour as in the patch models above.

The susceptible population in such spatio-temporal point process models consists of the continuous observation region and is thus infinite. The book chapter of Diggle (2007) also introduces other strategies for the analysis of spatial and spatio-temporal point patterns. An attractive alternative are multivariate temporal point process models for the event history of a fixed population. A corresponding endemic-epidemic formulation was proposed by Höhle (2009), which is also described and exemplified in Section 4 of Paper II. While statistical inference here is purely likelihood-based, Brown et al. (2014) applied such a model to aphid infestation on a sugar cane plantation, with a focus on Bayesian inference procedures to take interval-censored infection times into account. A class of *discrete-time* individual-level S-Exposed-IR models for a closed population is proposed by Deardon et al. (2010). Using a Bayesian framework as well, the model is applied to the 2001 UK foot-and-mouth epidemic extending the work of Keeling et al. (2001).

Yet unrelated to epidemic models, distinct methods exist to test for evidence of space-time clustering in point patterns. The Mantel (1967) test, for example, investigates the correlation betwen the pairwise spatial and temporal event distances. Such methods have been originally developed to study a possible dependence between cases of a disease with unknown etiology. These tests for space-time interaction have later also been taken up by other research communities, e.g., to analyse crime hot-spots (Grubesic and Mack, 2008). Paper IV proposes a more sophisticated model-based test approach by making use of the endemic-epidemic point process regression framework of Meyer et al. (2012).

Spatially explicit models for human diseases – and for epidemics in moving populations in general – suffer from a common issue: "Whether *place of residence*, which is how spatial position of individuals is usually measured, is a realistic assessment of where a person developed the disease or was exposed to the precipitating agent is often debatable" (Marshall, 1991). Similary, Euclidean or geodesic distance between individuals as illustrated in Figure 1b is a rather rough measure to quantify interaction. In spatially discrete models, such as those of Höhle (2009) and Deardon et al. (2010), Euclidean distance can be replaced by some "functional distance" (Brown and Horton, 1970) of (inter-)regional properties like population and commuter or travel flows (Brockmann and Helbing, 2013). Such models can then more generally be viewed as network rather than purely spatial epidemic models. Extensive reviews of network epidemic models are given by Danon et al. (2011) and Pastor-Satorras et al. (2015). In view of the growing amount of network data available (Salathé et al., 2012), "networks offer a fertile framework for studying the spread of infection in human and animal populations" (Pellis et al., 2015).

2 Thesis summary

This thesis consists of four papers and two appendices. Their content and contributions are summarised below.

Paper I: Power-law models for infectious disease spread

by Sebastian Meyer and Leonhard Held

Short-time human travel is dominated by local movements while longer journeys occur less frequently. More specifically, by studying the dispersal of dollar bills across the USA, Brockmann et al. (2006) found the distribution of travelling distances to decay as a power law. Since human travel is an important driver of epidemic spread, such a distance relation should enter spatio-temporal models for infectious disease occurrence. Paper I addresses this idea by embedding a power-law decay of spatial interaction in two previously established model classes: a spatio-temporal point process model for individual-level data (Meyer et al., 2012), and a multivariate time-series model for areal-level count data (Held and Paul, 2012). The paper reviews the model classes, discusses possible power-law formulations, and applies the extended models to public health surveillance data on invasive meningococcal disease and influenza. In both models, likelihood inference is carried out for all parameters simultaneously. Convergence is relatively fast, because the numerical maximization algorithm makes use of the analytically derived score function and Fisher information. The case studies reveal that the proposed power laws substantially improve model fit and predictions and are in line with alternative nonparametric formulations.

The idea to employ the power-law distribution of short-time human travel for more realistic transmission weights was originally mentioned in Held and Paul (2012), and formed the first research goal of this Ph.D. project. In discussions with L. Held, I investigated and implemented several power-law formulations for both model classes and worked out the corresponding inference machinery (detailed in Supplement B, Section 1). For the count data model, I generalised the existing implementation in the R package `surveillance` to allow for parametrised transmission weights. For the point process model, the new spatial interaction kernels additionally required more sophisticated numerical integration techniques over polygonal domains to compute the likelihood. I first translated a Matlab implementation of product Gauss cubature (Sommariva and Vianello, 2007) to R with portions in C. Since this method was not satisfactory in terms of accuracy and speed of the likelihood evaluations, I developed the idea for a more efficient cubature rule, which takes analytical advantage of the assumed isotropy of spatial interaction (the corresponding formula is derived in Supplement B, Section 2). I de-

cided to bundle these cubature methods in a dedicated R package `polyCub` (Meyer, 2016), which also enables others to use the implemented techniques (e.g., Quintero et al., 2015). I performed the two case studies and drafted the manuscript, on which L. Held commented. In particular, he suggested to compute additional long-term forecasts from the count data model and to evaluate these using proper scoring rules. I then finalised the manuscript and later incorporated helpful comments of M. Höhle, and two anonymous referees in the revised version. This included the additional estimation of nonparametric transmission weights to verify the power-law shape of spatial interaction. I have also investigated more flexible B-spline kernels in the point process model, but they were found to require a higher spatial resolution than typically available from public health surveillance systems (Appendix B). I have presented the proposed power-law models in talks at various conferences (DAGStat 2013, ROeS 2013, and SMIDDY 2013).

The main contribution of this paper are the proposed parsimonious spatial power-law formulations to model travel-induced infectious disease dynamics. Such spatial characteristics of disease spread can be estimated from the surveillance data jointly with all other model parameters. This is especially useful if additional movement network data are unavailable or error-prone, as concluded in a related project on modelling the spread of influenza in Northern Norway (Geilhufe et al., 2014). Both their research as well as a recent study in veterinary syndromic surveillance (Vial, Thommen and Held, 2015) were enabled by the ongoing development of the R package `surveillance`, which is described in Paper II.

Paper II: Spatio-temporal analysis of epidemic phenomena using the R package `surveillance`

by Sebastian Meyer, Leonhard Held, and Michael Höhle

The purpose of this paper is to document the rich set of features now readily available in the R package `surveillance` to analyse spatio-temporal epidemic data. In its three main parts, the paper illustrates the implementations of the two model classes from Paper I and the related multivariate point process model of Höhle (2009). Each of these models is suitable for a particular type of surveillance data and is exemplified in a case study with model building, estimation, diagnostics, simulation, and visualization. The unifying feature of the models is an additive decomposition of disease risk into endemic and epidemic components, which both include linear predictors of covariates similar to standard regression models. The role of exogeneous factors in shaping endemic and epidemic dynamics can thus be assessed. Typical examples of covariates are socio-economic characteristics, environmental conditions, and vaccination coverage. All models include measures of space-time interaction to be estimated from the data.

The reproducibility of research and the open-source implementation of statistical methods belong to my own core policies – especially considering that my Ph.D. project was publicly funded. However, the idea to write this paper for the Journal of Statistical Software is due to M. Höhle, the original maintainer of the `surveillance` package. I drafted the manuscript while simultaneously improving upon the usability of the package. M. Höhle contributed large parts of Section 4 as well as the first paragraph of Section 2. He commented on my drafts of the other sections, and L. Held commented on Sections 1, 2, 5 and 6. For the revision, I added a software review, and re-structured and extended the methodological descriptions in the manuscript. I also implemented and validated the calibration tests for count data of Wei and Held (2014) for inclusion in the `surveillance` package (based on original code of W. Wei). The complete set of changes to the package made in the context of this thesis is documented in the `NEWS` file available from the package's CRAN page (`https://CRAN.R-project.org/package=surveillance`) It also holds a copy of the comprehensive software manual. I have given a talk about this work at the useR! 2015 conference.

The main contribution of this paper is to make three recently established spatio-temporal modelling frameworks and their comprehensive implementation accessible to other researchers analysing epidemic data. This is especially important since "the modelling is clearly very complex" (an anonymous referee), and regression-oriented approaches to the spatio-temporal analysis of epidemic phenomena are currently not available in any other software package. Furthermore, some researchers in infectious disease epidemiology do not seem to be aware of the modelling approaches readily usable through the `surveillance` package and strive to reinvent the wheel (Imai et al., 2015). While the focus of the paper is on infectious disease epidemiology, the methodology similarly applies to other epidemic phenomena, including invasive species and crimes. For example, the `surveillance` package is currently being used by several research groups to analyse hand-foot-and-mouth disease, dengue fever, cholera, ranavirus-associated frog mortality, gun violence victimisation, and earthquakes.

Paper III: Incorporating social contact data in spatio-temporal models for infectious disease spread

by Sebastian Meyer and Leonhard Held

How likely people interact is not only characterised by their relative distance (Paper I), but also by demographic factors such as age and gender. An epidemiologically relevant form of interaction is face-to-face conversation, since it often defines an at-risk contact for pathogen transmission (Edmunds et al., 1997; Wallinga et al., 2006). However, conversational contacts are "highly assortative with age",

as quantified in a large social contact survey ("POLYMOD") across eight European countries (Mossong et al., 2008). For instance, school children mainly contact other children of the same age as well as their parents. Useful epidemic models should therefore take age structure into account – in addition to the spatial dynamics of disease spread. Paper III seeks to fill this gap by incorporating social contact data in an age-stratified version of the areal count time-series model from Paper I. For this purpose, the counts become indexed in three dimensions – age group, region, and time period – and accompanied by an empirically derived social contact matrix quantifying the interaction between age groups. The basic idea of this model extension is to superpose the contact rates between age groups (in a possibly adjusted form) as weights on top of the spatial power law. The paper exemplifies this procedure using routinely available public health surveillance data on norovirus gastroenteritis and an age-structured contact matrix estimated from the POLYMOD data. AIC-based model comparisons suggest that the incorporated contact matrix clearly outperforms naive models assuming homogeneous or no mixing between age groups. Furthermore, such a joint model formulation for the multivariate time series borrows strength across regions and age groups, which makes parameter estimates more efficient than in separate spatio-temporal models for each age group.

The idea to extend the spatio-temporal modelling framework of Paper I towards a unified analysis across age groups and regions was mentioned as an outlook in Held and Paul (2012). L. Held pointed me to the POLYMOD survey and suggested to somehow integrate such an empirically derived social contact matrix into the existing endemic-epidemic modelling framework. I proposed the age-stratified model formulation and extended the implementation in the `surveillance` package to allow for an additional transmission matrix as well as shared overdispersion parameters across groups of units. I gathered functionality specific to the new age-structured modelling in a dedicated R package `hhh4contacts`, which serves as an add-on to `surveillance` and is supplementary to the paper. Apart from implementing group-specific spatial transmission weights, profile likelihood estimation of the contact matrix adjustment, and plot methods for age-structured models, it also contains the contact and norovirus data used in the case study for it to be fully reproducible. While writing the manuscript, I discussed several model adaptions with L. Held. I have presented this work in talks at various conferences (61st Biometrical Congress 2015 of IBS-DR, IROeS 2015, and GEOMED 2015).

The main contribution of this paper is the unified modelling of disease spread across regions and population strata within regions, while incorporating external knowledge about contact rates between these strata. To the best of my knowledge, no other stratified spatio-temporal models for surveillance data are available at the time of this writing. However, Bauer and Wakefield (2015) are currently de-

veloping a similar extension of the areal count time-series model from Paper I to allow for stratification with respect to age and gender. Their model formulation differs in that the epidemic component retains the decomposition into local reproduction of the disease and imports from other regions. Furthermore, a Bayesian inference procedure is adopted, and, most importantly, social contact data are not involved. As an outlook from Paper III, L. Held and I investigated the predictive performance of the new age-structured model, which L. Held then presented at the Armitage Lecture 2015.[1]

Paper IV: Model-based testing for space-time interaction using point processes: An application to psychiatric hospital admissions in an urban area

by Sebastian Meyer, Ingeborg Warnke, Wulf Rössler, and Leonhard Held

This paper addresses the question of whether an observed spatio-temporal point pattern actually exhibits epidemic features. While this is natural for infectious diseases, tests for space-time interaction are of particular interest for diseases with unknown etiology. Several classical tests exist, which essentially evaluate the correlation between the spatial and temporal distances of the cases. However, these classical tests can be biased since they do not take spatio-temporally varying incidence into account and they do not allow the extent of clustering to be quantified. Paper IV proposes to employ the endemic-epidemic point process regression model of Paper I to estimate both a background incidence explained by the underlying population and a superposed infectious component representing the hypothesis of interest.

This work originates from a statistical consulting project for researchers from the psychiatric university hospital of Zurich (I. Warnke and W. Rössler). The specific question at hand is whether a psychiatric inpatient admission tends to trigger further admissions from the patient's neighbourhood – adjusting for regional socio-economic characteristics. It is hypothesised that each admission lowers help-seeking barriers in the local neighbourhood of the patient's residence, which eventually results in spatio-temporal clusters of inpatient admissions. This idea is related to models for the spread of rumours (Daley and Gani, 1999, Chapter 5). After having conducted an initial analysis with a classical test for space-time interaction, L. Held approached me with the idea of applying my work on point process models to address the research question in greater detail. I. Warnke provided all the necessary data for the modelling including the patient records as well as socio-economic characteristics of Zurich's quarters and a shapefile of the administrative city boundaries. She also wrote the application for ethical approval, for which I

[1] A videotaping with slides is available at http://view6.workcast.net/?pak=3248968452374653.

contributed the statistical parts. I proposed a permutation test to assess the evidence for an epidemic component and drafted the manuscript, on which L. Held, I. Warnke and M. Höhle commented. I applied several classical tests and the new model-based approach to the psychiatric inpatient data – implementing the necessary functionality in the `surveillance` package. For comparison, I also applied these tests to the data on invasive meningococcal disease already known from Paper I and included this analysis as supplementary material.

The main contribution of this paper is the proposal of a new global test for space-time interaction based on an endemic-epidemic point process regression model. This allows the test to be adjusted for covariates associated with the background rate of events, and to estimate and account for a distance decay of interaction.

Appendix A: A space-time conditional intensity model for invasive meningococcal disease occurrence

by Sebastian Meyer, Johannes Elias, and Michael Höhle

This paper has emerged from a technical report (Meyer et al., 2010) written before the start of my Ph.D. project. It summarises and extends the spatio-temporal point process model originally developed in my Master's Thesis (Meyer, 2009) supervised by M. Höhle. The paper is included here, since it lays the methodological foundations for several other parts of this thesis, including Paper I and Paper IV summarised above. A rich implementation of this model class is provided in the R package `surveillance` as described in Paper II (Section 3).

The main contribution of this paper is the formulation, implementation and application of a general modelling and inference framework for self-exciting spatio-temporal point processes. The model is specified in terms of the conditional intensity function of the process, which is additively decomposed into an endemic and an epidemic component. Both components incorporate log-linear predictors containing covariates associated with the endemic rate and the individual force of infection, respectively. The epidemic component additionally depends on parametric spatial and temporal interaction functions describing the distance decay and time course of infectivity, respectively. This novel spatio-temporal point process model class has been motivated by geo-referenced case reports of invasive meningococcal disease collected by the German Reference Centre for Meningococci (J. Elias). The selected model estimated the meningococcal finetype of serogroup B to be twice as infectious as the serogroup C type. There was no evidence for the time-lagged, local influenza incidence to increase the rate of cases of meningococcal disease. Furthermore, simulations from the fitted model helped to identify regions with an excess risk due to transmission across the border (edge effect).

Appendix B: Flexible estimation of spatio-temporal interaction in a point process model for infectious disease spread

by Sebastian Meyer and Leonhard Held

This short report follows up on Appendix A and Paper I with respect to obtaining more flexible estimates of spatial and temporal interaction in the aforementioned point process model. In Paper I, the proposed power-law decay of spatial interaction was checked against an unconstrained step function. While this is a computationally attractive alternative to parametric kernels, a drawback of the step function estimate is its strong dependence on the chosen knots, which also introduce unrealistic break points. In principle, higher-order B-splines may deliver additional insight into the spatial dependence structure or the time course of infectivity, respectively. However, they require the resolution of the point pattern to be high enough to identify the kernel shape over the whole range of event distances relevant for disease transmission. Furthermore, B-spline kernels considerably increase the computation time to evaluate the point process likelihood. I thus conclude that only the efficiently implemented step function kernel, i.e., a 0-degree B-spline, is useful as a quick benchmark of interaction. Otherwise, models should rather incorporate (empirically derived) parametric kernels, such as the Gaussian, Student, or power-law kernels for spatial interaction.

The idea of estimating a distance decay of interaction without relying on parametric assumptions on the shape of the kernel goes back to a comment of an anonymous reviewer of Paper I. After implementing the step function kernel for that paper, I investigated the use of higher-order B-splines. L. Held proposed linear basis functions on the log-log-scale, where the power law turns into a simple linear relation. He commented on my draft of this short report, which I then finalised. I had also briefly discussed the choice and integration of basis functions with M. Höhle, who suggested to look into alternative splines, which naturally ensure positivity. In this context, I would like to point to a flexible modelling option, for which I was only recently inspired by a talk of T. Hothorn: Bernstein polynomials (Farouki, 2012). These could be especially useful as spatial interaction functions, since monotonicity constraints can be easily handled.

References

Aldrin, M., Storvik, B., Kristoffersen, A. B. and Jansen, P. A. (2013). Space-time modelling of the spread of salmon lice between and within Norwegian marine salmon farms, *PLOS ONE* **8**(5): e64039. doi:10.1371/journal.pone.0064039.

Anderson, R. M. and May, R. M. (1991). *Infectious Diseases of Humans: Dynamics and Control*, Oxford University Press.

Andersson, E., Bock, D. and Frisén, M. (2008). Modeling influenza incidence for the purpose of on-line monitoring, *Statistical Methods In Medical Research* **17**(4): 421–438. doi:10.1177/0962280206078986.

Andersson, H. and Britton, T. (2000). *Stochastic Epidemic Models and Their Statistical Analysis*, Vol. 151 of *Lecture Notes in Statistics*, Springer-Verlag, New York.

Balcan, D., Colizza, V., Gonçalves, B., Hu, H., Ramasco, J. J. and Vespignani, A. (2009). Multiscale mobility networks and the spatial spreading of infectious diseases, *Proceedings of the National Academy of Sciences of the United States of America* **106**(51): 21484–21489. doi:10.1073/pnas.0906910106.

Balderama, E., Schoenberg, F. P., Murray, E. and Rundel, P. W. (2012). Application of branching models in the study of invasive species, *Journal of the American Statistical Association* **107**(498): 467–476. doi:10.1080/01621459.2011.641402.

Ball, F., Britton, T., House, T., Isham, V., Mollison, D., Pellis, L. and Tomba, G. S. (2015). Seven challenges for metapopulation models of epidemics, including households models, *Epidemics* **10**: 63–67. doi:10.1016/j.epidem.2014.08.001.

Bauer, C. and Wakefield, J. (2015). Stratified space-time infectious disease modeling: with an application to hand, foot and mouth disease in China. In preparation.

Bauer, C., Wakefield, J., Rue, H., Self, S., Feng, Z. and Wang, Y. (2016). Bayesian penalized spline models for the analysis of spatio-temporal count data, *Statistics in Medicine* **35**(11): 1848–1865. doi:10.1002/sim.6785.

Becker, N. G. (1989). *Analysis of Infectious Disease Data*, Monographs on Statistics and Applied Probability, Chapman & Hall, London.

Becker, N. G. and Britton, T. (1999). Statistical studies of infectious disease incidence, *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **61**(2): 287–307. doi:10.1111/1467-9868.00177.

Britton, T. (2010). Stochastic epidemic models: A survey, *Mathematical Biosciences* **225**(1): 24–35. doi:10.1016/j.mbs.2010.01.006.

Brockmann, D. and Helbing, D. (2013). The hidden geometry of complex, network-driven contagion phenomena, *Science* **342**(6164): 1337–1342. doi:10.1126/science.1245200.

Brockmann, D., Hufnagel, L. and Geisel, T. (2006). The scaling laws of human travel, *Nature* **439**(7075): 462–465. doi:10.1038/nature04292.

Brooks-Pollock, E., de Jong, M., Keeling, M., Klinkenberg, D. and Wood, J. (2015). Eight challenges in modelling infectious livestock diseases, *Epidemics* **10**: 1–5. doi:10.1016/j.epidem.2014.08.005.

Brown, L. A. and Horton, F. E. (1970). Functional distance: an operational approach, *Geographical Analysis* **2**(1): 76–83. doi:10.1111/j.1538-4632.1970.tb00146.x.

Brown, P. E., Chimard, F., Remorov, A., Rosenthal, J. S. and Wang, X. (2014). Statistical inference and computational efficiency for spatial infectious disease models with plantation data, *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **63**(3): 467–482. doi:10.1111/rssc.12036.

Cauchemez, S., Bhattarai, A., Marchbanks, T. L., Fagan, R. P., Ostroff, S., Ferguson, N. M., Swerdlow, D. and the Pennsylvania H1N1 working group (2011). Role of social networks in shaping disease transmission during a community outbreak of 2009 H1N1 pandemic influenza, *Proceedings of the National Academy of Sciences of the United States of America* **108**(7): 2825–2830. doi:10.1073/pnas.1008895108.

Chretien, J.-P., George, D., Shaman, J., Chitale, R. A. and McKenzie, F. E. (2014). Influenza forecasting in human populations: A scoping review, *PLOS ONE* **9**(4): e94130. doi:10.1371/journal.pone.0094130.

Cunniffe, N. J., Koskella, B., Metcalf, C. J. E., Parnell, S., Gottwald, T. R. and Gilligan, C. A. (2015). Thirteen challenges in modelling plant diseases, *Epidemics* **10**: 6–10. doi:10.1016/j.epidem.2014.06.002.

Daley, D. J. and Gani, J. (1999). *Epidemic Modelling: An Introduction*, Cambridge University Press.

Danon, L., Ford, A. P., House, T., Jewell, C. P., Keeling, M. J., Roberts, G. O., Ross, J. V. and Vernon, M. C. (2011). Networks and the epidemiology of infectious disease, *Interdisciplinary Perspectives on Infectious Diseases* **2011**: 284909. doi:10.1155/2011/284909.

Deardon, R., Brooks, S., Grenfell, B., Keeling, M., Tildesley, M., Savill, N., Shaw, D. and Woolhouse, M. (2010). Inference for individual-level models of infectious diseases in large populations, *Statistica Sinica* **20**(1): 239–261.

Deardon, R., Fang, X. and Kwong, G. P. S. (2015). Statistical Modeling of Spatiotemporal Infectious Disease Transmission, *in* D. Chen, B. Moulin and J. Wu (eds), *Analyzing and Modeling Spatial and Temporal Dynamics of Infectious Diseases*, Wiley Series in Probability and Statistics, John Wiley & Sons, chapter 11, pp. 211–231. doi:10.1002/9781118630013.ch11.

Diggle, P. J. (2007). Spatio-Temporal Point Processes: Methods and Applications, *in* B. Finkenstädt, L. Held and V. Isham (eds), *Statistical Methods for Spatio-Temporal Systems*, Chapman & Hall/CRC, Boca Raton, chapter 1, pp. 1–45.

Diggle, P. J., Rowlingson, B. and Su, T.-l. (2005). Point process methodology for on-line spatio-temporal disease surveillance, *Environmetrics* **16**(5): 423–434. doi:10.1002/env.712.

Eames, K., Bansal, S., Frost, S. and Riley, S. (2015). Six challenges in measuring contact networks for use in modelling, *Epidemics* **10**: 72–77. doi:10.1016/j.epidem.2014.08.006.

Edmunds, W. J., O'Callaghan, C. J. and Nokes, D. J. (1997). Who mixes with whom? A method to determine the contact patterns of adults that may lead to the spread of airborne infections, *Proceedings of the Royal Society of London B: Biological Sciences* **264**(1384): 949–957. doi:10.1098/rspb.1997.0131.

Eggo, R. M., Cauchemez, S. and Ferguson, N. M. (2011). Spatial dynamics of the 1918 influenza pandemic in England, Wales and the United States, *Journal of the Royal Society Interface* **8**(55): 233–243. doi:10.1098/rsif.2010.0216.

Ensoy, C., Aerts, M., Welby, S., Van der Stede, Y. and Faes, C. (2013). A dynamic spatio-temporal model to investigate the effect of cattle movements on the spread of bluetongue BTV-8 in Belgium, *PLOS ONE* **8**(11): e78591. doi:10.1371/journal.pone.0078591.

Faensen, D. and Krause, G. (2004). SurvStat@RKI – a web-based solution to query surveillance data in Germany, *Eurosurveillance* **8**(22). URL: http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=2477.

Farouki, R. T. (2012). The Bernstein polynomial basis: A centennial retrospective, *Computer Aided Geometric Design* **29**(6): 379–419. doi:10.1016/j.cagd.2012.03.001.

Ferguson, N. M., Donnelly, C. A. and Anderson, R. M. (2001). The foot-and-mouth epidemic in Great Britain: Pattern of spread and impact of interventions, *Science* **292**(5519): 1155–1160. doi:10.1126/science.1061020.

Finkenstädt, B. F., Bjørnstad, O. N. and Grenfell, B. T. (2002). A stochastic model for extinction and recurrence of epidemics: estimation and inference for measles outbreaks, *Biostatistics* **3**(4): 493–510. doi:10.1093/biostatistics/3.4.493.

Finkenstädt, B. F. and Grenfell, B. T. (2000). Time series modelling of childhood diseases: a dynamical systems approach, *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **49**(2): 187–205. doi:10.1111/1467-9876.00187.

Fèvre, E. M., de C. Bronsvoort, B. M., Hamilton, K. A. and Cleaveland, S. (2006). Animal movements and the spread of infectious diseases, *Trends in Microbiology* **14**(3): 125–131. doi:10.1016/j.tim.2006.01.004.

Geilhufe, M., Held, L., Skrøvseth, S. O., Simonsen, G. S. and Godtliebsen, F. (2014). Power law approximations of movement network data for modeling infectious disease spread, *Biometrical Journal* **56**(3): 363–382. doi:10.1002/bimj.201200262.

Giesecke, J. (2002). *Modern Infectious Disease Epidemiology*, 2nd edn, Arnold, London, United Kingdom.

Gneiting, T. and Katzfuss, M. (2014). Probabilistic forecasting, *Annual Review of Statistics and Its Application* **1**(1): 125–151. doi:10.1146/annurev-statistics-062713-085831.

Goeyvaerts, N., Hens, N., Ogunjimi, B., Aerts, M., Shkedy, Z., van Damme, P. and Beutels, P. (2010). Estimating infectious disease parameters from data on social contacts and serological status, *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **59**(2): 255–277. doi:10.1111/j.1467-9876.2009.00693.x.

Gog, J. R., Ballesteros, S., Viboud, C., Simonsen, L., Bjørnstad, O. N., Shaman, J., Chao, D. L., Khan, F. and Grenfell, B. T. (2014). Spatial transmission of 2009 pandemic influenza in the US, *PLOS Computational Biology* **10**(6): e1003635. doi:10.1371/journal.pcbi.1003635.

Grubesic, T. H. and Mack, E. A. (2008). Spatio-temporal interaction of urban crime, *Journal of Quantitative Criminology* **24**(3): 285–306. doi:10.1007/s10940-008-9047-5.

Held, L., Höhle, M. and Hofmann, M. (2005). A statistical framework for the analysis of multivariate infectious disease surveillance counts, *Statistical Modelling* **5**(3): 187–199. doi:10.1191/1471082X05st098oa.

Held, L. and Paul, M. (2012). Modeling seasonality in space-time infectious disease surveillance data, *Biometrical Journal* **54**(6): 824–843. doi:10.1002/bimj.201200037.

Held, L. and Paul, M. (2013). Statistical modeling of infectious disease surveillance data, *in* N. M. M'ikanatha, R. Lynfield, C. A. V. Beneden and H. de Valk (eds), *Infectious Disease Surveillance*, 2nd edn, John Wiley & Sons, chapter 36, pp. 535–544. doi:10.1002/9781118543504.ch43.

Hens, N., Shkedy, Z., Aerts, M., Faes, C., Van Damme, P. and Beutels, P. (2012). *Modeling Infectious Disease Parameters Based on Serological and Social Contact Data*, Vol. 63 of *Statistics for Biology and Health*, Springer, New York. doi:10.1007/978-1-4614-4072-7.

Höhle, M. (2009). Additive-multiplicative regression models for spatio-temporal epidemics, *Biometrical Journal* **51**(6): 961–978. doi:10.1002/bimj.200900050.

Höhle, M. (2016). Infectious Disease Modelling, *in* A. B. Lawson, S. Banerjee, R. P. Haining and M. D. Ugarte (eds), *Handbook of Spatial Epidemiology*, Chapman & Hall/CRC Handbooks of Modern Statistical Methods, Chapman and Hall/CRC, chapter 26, pp. 477–500.

Höhle, M., Meyer, S. and Paul, M. (2016). *surveillance: Temporal and Spatio-Temporal Modeling and Monitoring of Epidemic Phenomena*. R package version 1.12.1. URL: http://surveillance.r-forge.r-project.org/.

Imai, C., Armstrong, B., Chalabi, Z., Mangtani, P. and Hashizume, M. (2015). Time series regression model for infectious disease and weather, *Environmental Research* **142**: 319–327. doi:10.1016/j.envres.2015.06.040.

Keeling, M. J., Bjørnstad, O. N. and Grenfell, B. T. (2004). Metapopulation Dynamics of Infectious Diseases, *in* I. Hanski and O. E. Gaggiotti (eds), *Ecology, Genetics and Evolution of Metapopulations*, Academic Press, Burlington, chapter 17, pp. 415–445. doi:10.1016/B978-012323448-3/50019-2.

Keeling, M. J. and Rohani, P. (2008). *Modeling Infectious Diseases in Humans and Animals*, Princeton University Press. URL: http://www.modelinginfectiousdiseases.org/.

Keeling, M. J., Woolhouse, M. E. J., Shaw, D. J., Matthews, L., Chase-Topping, M., Haydon, D. T., Cornell, S. J., Kappey, J., Wilesmith, J. and Grenfell, B. T. (2001). Dynamics of the 2001 UK foot and mouth epidemic: stochastic dispersal in a heterogeneous landscape, *Science* **294**(5543): 813–817. doi:10.1126/science.1065973.

Kermack, W. O. and McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics, *Proceedings of the Royal Society of London. Series A* **115**(772): 700–721. doi:10.1098/rspa.1927.0118.

Lawson, A. B. and Kleinman, K. (eds) (2005). *Spatial and Syndromic Surveillance for Public Health*, Wiley. doi:10.1002/0470092505.

Lloyd-Smith, J. O., Mollison, D., Metcalf, C. J. E., Klepac, P. and Heesterbeek, J. (2015). Challenges in Modelling Infectious Disease Dynamics: Preface, *Epidemics* **10**: iii–iv. doi:10.1016/j.epidem.2015.02.001.

Madden, L. V., Hughes, G. and van den Bosch, F. (2007). *The Study of Plant Disease Epidemics*, APS Press.

Manitz, J. and Höhle, M. (2013). Bayesian outbreak detection algorithm for monitoring reported cases of campylobacteriosis in Germany, *Biometrical Journal* **55**(4): 509–526. doi:10.1002/bimj.201200141.

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach, *Cancer Research* **27**(2): 209–220. URL: http://cancerres.aacrjournals.org/content/27/2_Part_1/209.abstract.

Marshall, R. J. (1991). A review of methods for the statistical analysis of spatial patterns of disease, *Journal of the Royal Statistical Society. Series A (Statistics in Society)* **154**(3): 421–441. URL: http://www.jstor.org/stable/2983152.

Meyer, S. (2009). *Spatio-Temporal Infectious Disease Epidemiology based on Point Processes*, Master's thesis, Department of Statistics, Ludwig-Maximilians-Universität, München. URL: http://epub.ub.uni-muenchen.de/11703/.

Meyer, S. (2016). *polyCub: Cubature over Polygonal Domains*. R package version 0.5-2. URL: https://CRAN.R-project.org/package=polyCub.

Meyer, S., Elias, J. and Höhle, M. (2010). A space-time conditional intensity model for infectious disease occurence, *Technical Report 95*, Department of Statistics, Ludwig-Maximilians-Universität München. URL: http://epub.ub.uni-muenchen.de/11898/.

Meyer, S., Elias, J. and Höhle, M. (2012). A space-time conditional intensity model for invasive meningococcal disease occurrence, *Biometrics* **68**(2): 607–616. URL: https://epub.ub.uni-muenchen.de/25195/, arXiv:1508.05740, doi:10.1111/j.1541-0420.2011.01684.x.

Mohler, G. O., Short, M. B., Malinowski, S., Johnson, M., Tita, G. E., Bertozzi, A. L. and Brantingham, P. J. (2015). Randomized controlled field trials of predictive policing, *Journal of the American Statistical Association* **110**(512): 1399–1411. doi:10.1080/01621459.2015.1077710.

Mossong, J., Hens, N., Jit, M., Beutels, P., Auranen, K., Mikolajczyk, R., Massari, M., Salmaso, S., Tomba, G. S., Wallinga, J., Heijne, J., Sadkowska-Todys, M., Rosinska, M. and Edmunds, W. J. (2008). Social contacts and mixing patterns relevant to the spread of infectious diseases, *PLoS Medicine* **5**(3): e74. doi:10.1371/journal.pmed.0050074.

Ogata, Y. (1998). Space-time point-process models for earthquake occurrences, *Annals of the Institute of Statistical Mathematics* **50**(2): 379–402. doi:10.1023/A:1003403601725.

O'Neill, P. D. (2010). Introduction and snapshot review: relating infectious disease transmission models to data., *Statistics in Medicine* **29**(20): 2069–2077. doi:10.1002/sim.3968.

Pastor-Satorras, R., Castellano, C., Van Mieghem, P. and Vespignani, A. (2015). Epidemic processes in complex networks, *Reviews of Modern Physics* **87**(3): 925–979. doi:10.1103/RevModPhys.87.925.

Paul, M. and Held, L. (2011). Predictive assessment of a non-linear random effects model for multivariate time series of infectious disease counts, *Statistics in Medicine* **30**(10): 1118–1136. doi:10.1002/sim.4177.

Paul, M., Held, L. and Toschke, A. (2008). Multivariate modelling of infectious disease surveillance data, *Statistics in Medicine* **27**(29): 6250–6267. doi:10.1002/sim.3440.

Pellis, L., Ball, F., Bansal, S., Eames, K., House, T., Isham, V. and Trapman, P. (2015). Eight challenges for network epidemic models, *Epidemics* **10**: 58–62. doi:10.1016/j.epidem.2014.07.003.

Piroutek, A., Assunção, R. and Paiva, T. (2014). Space-time prospective surveillance based on Knox local statistics, *Statistics in Medicine* **33**(16): 2758–2773. doi:10.1002/sim.6118.

Quintero, I., Keil, P., Jetz, W. and Crawford, F. W. (2015). Historical biogeography using species geographical ranges, *Systematic Biology* **64**(6): 1059–1073. doi:10.1093/sysbio/syv057.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

Riley, S. (2007). Large-scale spatial-transmission models of infectious disease, *Science* **316**(5829): 1298–1301. doi:10.1126/science.1134695.

Riley, S., Eames, K., Isham, V., Mollison, D. and Trapman, P. (2015). Five challenges for spatial epidemic models, *Epidemics* **10**: 68–71. doi:10.1016/j.epidem.2014.07.001.

Salathé, M., Bengtsson, L., Bodnar, T. J., Brewer, D. D., Brownstein, J. S., Buckee, C., Campbell, E. M., Cattuto, C., Khandelwal, S., Mabry, P. L. and Vespignani, A. (2012). Digital epidemiology, *PLoS Computational Biology* **8**(7): e1002616. doi:10.1371/journal.pcbi.1002616.

Salmon, M., Schumacher, D. and Höhle, M. (2016). Monitoring count time series in R: Aberration detection in public health surveillance, *Journal of Statistical Software* **70**(10): 1–35. arXiv:1411.1292, doi:10.18637/jss.v070.i10.

Schrödle, B., Held, L. and Rue, H. (2012). Assessing the impact of a movement network on the spatiotemporal spread of infectious diseases, *Biometrics* **68**(3): 736–744. doi:10.1111/j.1541-0420.2011.01717.x.

Sommariva, A. and Vianello, M. (2007). Product Gauss cubature over polygons based on Green's integration formula, *Bit Numerical Mathematics* **47**(2): 441–453. doi:10.1007/s10543-007-0131-2.

Sonesson, C. and Bock, D. (2003). A review and discussion of prospective statistical surveillance in public health, *Journal of the Royal Statistical Society. Series A (Statistics in Society)* **166**(1): 5–21. doi:10.1111/1467-985X.00256.

Unkel, S., Farrington, C. P., Garthwaite, P. H., Robertson, C. and Andrews, N. (2012). Statistical methods for the prospective detection of infectious disease outbreaks: a review, *Journal of the Royal Statistical Society. Series A (Statistics in Society)* **175**(1): 49–82. doi:10.1111/j.1467-985X.2011.00714.x.

Vial, F., Thommen, S. and Held, L. (2015). A simulation study on the statistical monitoring of condemnation rates from slaughterhouses for syndromic surveillance: an evaluation based on Swiss data, *Epidemiology & Infection* **143**(16): 3423–3433. doi:10.1017/S0950268815000989.

Vial, F., Wei, W. and Held, L. (2015). Multivariate syndromic surveillance: Methodological challenges with a focus on animal health data. Submitted.

Viboud, C., Bjørnstad, O. N., Smith, D. L., Simonsen, L., Miller, M. A. and Grenfell, B. T. (2006). Synchrony, waves, and spatial hierarchies in the spread of influenza, *Science* **312**(5772): 447–451. doi:10.1126/science.1125237.

Wallinga, J., Teunis, P. and Kretzschmar, M. (2006). Using data on social contacts to estimate age-specific transmission parameters for respiratory-spread infectious agents, *American Journal of Epidemiology* **164**(10): 936–944. doi:10.1093/aje/kwj317.

Wei, W. and Held, L. (2014). Calibration tests for count data, *Test* **23**(4): 787–805. doi:10.1007/s11749-014-0380-8.

WHO Ebola Response Team (2014). Ebola virus disease in West Africa — The first 9 months of the epidemic and forward projections, *New England Journal of Medicine* **371**(16): 1481–1495. doi:10.1056/NEJMoa1411100.

Xia, Y., Bjørnstad, O. N. and Grenfell, B. T. (2004). Measles metapopulation dynamics: A gravity model for epidemiological coupling and dynamics, *The American Naturalist* **164**(2): 267–281. URL: http://www.jstor.org/stable/10.1086/422341.

PAPER I

Power-law models for infectious disease spread

Sebastian Meyer, Leonhard Held

Published in *The Annals of Applied Statistics*, 2014, **8** (3), 1612–1638.

The Annals of Applied Statistics
2014, Vol. 8, No. 3, 1612–1639
DOI: 10.1214/14-AOAS743

POWER-LAW MODELS FOR INFECTIOUS DISEASE SPREAD[1]

BY SEBASTIAN MEYER AND LEONHARD HELD

University of Zurich

Short-time human travel behaviour can be described by a power law with respect to distance. We incorporate this information in space–time models for infectious disease surveillance data to better capture the dynamics of disease spread. Two previously established model classes are extended, which both decompose disease risk additively into endemic and epidemic components: a spatio-temporal point process model for individual-level data and a multivariate time-series model for aggregated count data. In both frameworks, a power-law decay of spatial interaction is embedded into the epidemic component and estimated jointly with all other unknown parameters using (penalised) likelihood inference. Whereas the power law can be based on Euclidean distance in the point process model, a novel formulation is proposed for count data where the power law depends on the order of the neighbourhood of discrete spatial units. The performance of the new approach is investigated by a reanalysis of individual cases of invasive meningococcal disease in Germany (2002–2008) and count data on influenza in 140 administrative districts of Southern Germany (2001–2008). In both applications, the power law substantially improves model fit and predictions, and is reasonably close to alternative qualitative formulations, where distance and order of neighbourhood, respectively, are treated as a factor. Implementation in the R package *surveillance* allows the approach to be applied in other settings.

1. Introduction. The surveillance of infectious diseases constitutes a key issue of public health and modelling their spread is basic to the prevention and control of epidemics. An important task is the timely detection of disease outbreaks, for which popular methods are the Farrington algorithm [Farrington et al. (1996), Noufaily et al. (2013)] and cumulative sum (CUSUM) likelihood ratio detectors inspired by statistical process control [Höhle and Paul (2008), Höhle, Paul and Held (2009)]. As opposed to such prospective surveillance, retrospective surveillance is concerned with explaining the spread of an epidemic through statistical modelling, thereby assessing the role of environmental and socio-demographic factors or contact networks in shaping the evolution of an epidemic. The spatio-temporal data for such modelling primarily originate from routine public health surveillance of the occurrence of infectious diseases and is ideally accompanied by additional data on

Received August 2013; revised February 2014.
[1]Funded by the Swiss National Science Foundation (project #137919).
Key words and phrases. Power law, spatial interaction function, infectious disease surveillance, stochastic epidemic modelling, branching process with immigration, multivariate time series of counts, spatio-temporal point process.

influential factors to be accounted for. Surveillance data are available in different spatio-temporal resolutions, each type requiring an appropriate model framework.

This paper covers both a spatio-temporal point process model for individual-level data [proposed by Meyer, Elias and Höhle (2012) and motivated by the work of Höhle (2009)] and a multivariate time-series model for aggregated count data [established by Held and Paul (2012) and earlier work]. Although these two models are designed for different types of spatio-temporal surveillance data, both are inspired by the approach of Held, Höhle and Hofmann (2005) decomposing disease risk additively into "endemic" and "epidemic" components. The endemic component captures exogenous factors such as population, socio-demographic variables, long-term trends, seasonality, climate, or concurrent incidence of related diseases (all varying in time and/or space). Explicit dependence between cases, that is, infectiousness, is then introduced through epidemic components driven by the observed past.

To describe disease spread in space, both models account for spatial interaction between units or individuals, respectively, but up to now, this has been incorporated rather crudely. The point process model used a Gaussian kernel to capture spatial interaction, and the multivariate time-series model restricted epidemic spread from time t to $t+1$ to adjacent regions. However, a simple form of dispersal can be motivated by the findings of Brockmann, Hufnagel and Geisel (2006): they inferred from the dispersal of bank notes in the United States that (short-time) human travel behaviour can be well described by a decreasing power law of the distance x, that is, $f(x) \propto x^{-d}$ with positive decay parameter d. An important characteristic of this power law is its slow convergence to zero ("heavy tail"), which in our application enables occasional long-range transmissions of infectious agents in addition to principal short-range infections. In the words of Brockmann, Hufnagel and Geisel (2006), their results "can serve as a starting point for the development of a new class of models for the spread of human infectious diseases". Power laws are well known from the work by Pareto (1896) for the distribution of income and Zipf (1949) for city sizes and word frequencies in texts. They describe the distribution of earthquake magnitudes [Gutenberg and Richter (1944)] and many other natural phenomena [see Newman (2005), Pinto, Mendes Lopes and Machado (2012), for a review of power laws]. Liljeros et al. (2001) reported on a power-law distribution of the number of sexual partners, and Albert and Barabási (2002) review recent advances in network theory including scale-free networks where the number of edges is distributed according to a power law. Interestingly, a power law was also used as the distance decay function in geographic profiling for serial violent crime investigation [Rossmo (2000)] as well as in an application of this technique to identify environmental sources of infection [Le Comber et al. (2011)]. Examples of power-law transmission kernels to model the spatial dynamics of infectious diseases can be found in plant epidemiology [Gibson (1997), Soubeyrand et al. (2008)] and in models for the 2001 UK foot-and-mouth disease epidemic [Chis Ster and Ferguson (2007)]. Recently, Geilhufe et al. (2014) found that using (fixed) power-law

weights between regions performed better than real traffic data in predicting influenza counts in Northern Norway. In both models for spatio-temporal surveillance data presented in the following sections, the power law will be estimated jointly with all other unknown parameters. Since the choice of a power law is a strong (yet well motivated) assumption, a comparison with alternative qualitative formulations is provided.

This paper is organised as follows: in Sections 2 and 3, respectively, the two model frameworks are reviewed and extended with power-law formulations for the spatial interaction of units. In Section 4 surveillance data on invasive meningococcal disease (IMD) and influenza are reanalysed using power laws and alternative qualitative approaches to be evaluated against previously used models for these data. We close with some discussion in Section 5 and a software overview in the Appendix. The paper is accompanied by animations (Supplement A) and further supplementary material [Supplement B: Meyer and Held (2014)].

2. Individual-level model.

2.1. *Introduction*. The spatio-temporal point process model proposed by Meyer, Elias and Höhle (2012) is designed for time–space-mark data $\{(t_i, \mathbf{s}_i, \mathbf{m}_i) : i = 1, \ldots, n\}$ of individual case reports to describe the occurrence of infections ('events') and their potential to trigger secondary cases. Formally, the model characterises a point process in a region $\mathbf{W}$ observed during a period $(0, T]$ through the conditional intensity function

$$\lambda(t, \mathbf{s}) = \nu_{[t][\mathbf{s}]}\rho_{[t][\mathbf{s}]} + \sum_{j : t_j < t} \eta_j \cdot g(t - t_j) \cdot f\big(\|\mathbf{s} - \mathbf{s}_j\|\big). \tag{1}$$

Related models are the purely temporal, "self-exciting" process proposed by Hawkes (1971), the spatio-temporal epidemic-type aftershock-sequences (ETAS) model from earthquake research [Ogata (1998)], the point process models discussed by Diggle (2007), and an additive-multiplicative point process model for discrete-space surveillance data proposed by Höhle (2009).

The first endemic component in model (1) consists of a log-linear predictor $\log(\nu_{[t][\mathbf{s}]}) = \beta_0 + \boldsymbol{\beta}^\top \mathbf{z}_{[t][\mathbf{s}]}$ proportional to an offset $\rho_{[t][\mathbf{s}]}$, typically the population density. Both the offset and the exogenous covariates are given piecewise constant on a spatio-temporal grid (e.g., week $\times$ district), hence the notation $[t][\mathbf{s}]$ for the period which contains t in the region covering $\mathbf{s}$. In the IMD application in Section 4.1, $\mathbf{z}_{[t][\mathbf{s}]} = ([t], \sin(\omega \cdot [t]), \cos(\omega \cdot [t]))^\top$ incorporates a time trend with one sinusoidal wave of frequency $\omega = 2\pi/365$.

A purely endemic intensity model without the observation-driven epidemic component is equivalent to a Poisson regression model for the aggregated number of cases on the chosen spatio-temporal grid. However, with an epidemic component the intensity process depends on previously infected individuals and becomes "self-exciting." Specifically, the epidemic force of infection at $(t, \mathbf{s})$ is

the superposition of the infection pressures caused by each previously infected individual j. The individual infection pressure is weighted by the log-linear predictor $\log(\eta_j) = \gamma_0 + \boldsymbol{\gamma}^\top \mathbf{m}_j$, which models effects of individual/infection-specific characteristics $\mathbf{m}_j$ such as the age of the infective. Regional-level covariates could also be included in $\mathbf{m}_j$, for example, to model ecological effects on infectivity. Note, however, since the epidemic is modelled through a point process, the susceptible "population" consists of the continuous observation region $\mathbf{W} \subset \mathbb{R}^2$ and is thus infinite. Consequently, the model cannot include information on susceptibles, nor an autoregressive term as in time-series models.

Decreasing infection pressure of individual j over space and time is described by $f(x)$ and $g(t)$, parametric functions of the spatial distance x and of the elapsed time t since individual j became infectious, respectively. The spatial interaction could also be described more generally by a nonisotropic function $f_2(\mathbf{s})$ of the vector $\mathbf{s}$ to the host, for example, to incorporate the dominant wind direction in vector-borne diseases. However, in our application, f essentially reflects people's movements and we assume that $f_2(\mathbf{s}) = f(\|\mathbf{s}\|)$ only depends on the distance to the host. Note that we project geographic coordinates into a planar coordinate reference system to apply Euclidean geometry. Meyer, Elias and Höhle (2012) used an isotropic Gaussian kernel

$$f(x) = \exp\left(-\frac{x^2}{2\sigma^2}\right) \tag{2}$$

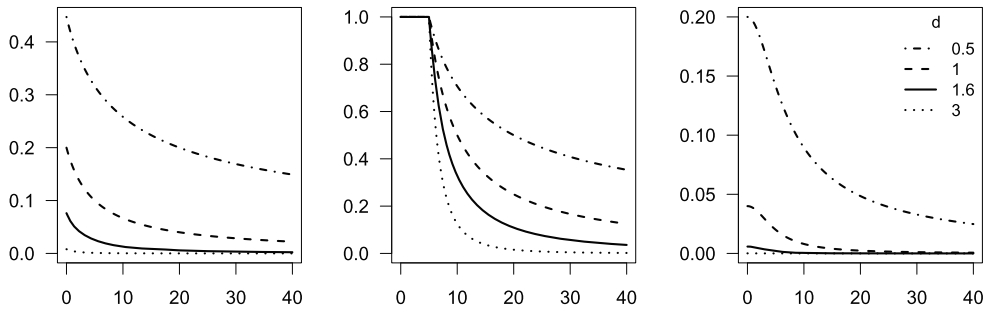
with scale parameter σ. In what follows, we propose an alternative spatial interaction function, which allows for occasional long-range transmission of infections: a power law.

2.2. *Power-law extension.* The basic power law $f(x) = x^{-d}$, $d > 0$, is not a suitable choice for the distance decay of infectivity since it has a pole at $x = 0$. For $x \geq \sigma > 0$, x^{-d} is the kernel of a Pareto density, but a shifted version to the domain $\mathbb{R}_0^+$, known as Pareto type II and sometimes named after Lomax (1954), has density kernel

$$f(x) = (x + \sigma)^{-d} \propto \left(1 + \frac{x}{\sigma}\right)^{-d} \tag{3}$$

[see Figure 1(a)]. Note that there is no need for the spatial interaction function to be normalised to a density. It is actually more closely related to correlation functions known from stationary random field models for geostatistical data [Chilès and Delfiner (2012)]. For instance, the rescaled version $(1 + x/\sigma)^{-d}$ is a member of the Cauchy class introduced by Gneiting and Schlather (2004), which provides asymptotic power-law correlation as $x \to \infty$.

For short-range travel within 10 km, Brockmann, Hufnagel and Geisel (2006) found a uniform distribution instead of power-law behaviour, which suggests an



(a) Power-law kernel (b) Lagged power law (c) Student kernel

FIG. 1. *Power-law kernels as a function of the distance x for various choices of the decay parameter $d > 0$ and $\sigma = 5$.*

alternative formulation with a "lagged" power law:

$$f(x) = \begin{cases} 1, & \text{for } x \leq \sigma, \\ \left(\dfrac{x}{\sigma}\right)^{-d}, & \text{otherwise.} \end{cases} \tag{4}$$

Spatial interaction is now constant up to the change point $\sigma > 0$, followed by a power-law decay for larger distances [see Figure 1(b)]. A similar kernel was used by Deardon et al. [(2010), therein called "geometric"] for the 2001 UK foot-and-mouth disease epidemic, additionally limiting spatial interaction to a prespecified upper-bound distance.

A change-point-free kernel also unifying intended short-range and long-range characteristics is the Student kernel

$$f(x) = \left(x^2 + \sigma^2\right)^{-d} \propto \left(1 + \left(\frac{x}{\sigma}\right)^2\right)^{-d} \tag{5}$$

with scale parameter σ and shape (decay) parameter d [see Figure 1(c)]. This kernel implements a power law of the squared distance and is known as the 'Cauchy model' in geostatistics [Chilès and Delfiner (2012)]. For $d > 0.5$, it describes a Student distribution with $2d - 1$ degrees of freedom.

To investigate the appropriateness of the assumed power-law decay, we also estimate an unconstrained step function

$$f(x) = \sum_{k=0}^{K} \alpha_k \mathbb{1}(x \in I_k), \tag{6}$$

which corresponds to treating the distance x—categorised into consecutive intervals I_k—as a qualitative variable.

2.3. *Inference.* Model parameters are estimated via maximization of the full (log-)likelihood, applying a quasi-Newton algorithm with analytical gradient and

Hessian [see Meyer and Held (2014), Section 1.1]. We estimate kernel parameters on the log-scale to avoid constrained optimization. For the step function, $\alpha_0 = 1$ is fixed to ensure identifiability.

The point process likelihood incorporates the integral of $f_2(\mathbf{s})$ over shifted versions of the observation region $\mathbf{W}$, which is represented by polygons. Similar integrals arise for the partial derivatives of $f_2(\boldsymbol{s})$ in the score function and approximate Fisher information. Except for the step function kernel (6), this requires a method of numerical integration such as the two-dimensional midpoint rule with an adaptive bandwidth, which was found to be best suited for the Gaussian kernel [Meyer, Elias and Höhle (2012)]. For the other kernels we use a more sophisticated approach inspired by product Gauss cubature over polygons [Sommariva and Vianello (2007)]. This cubature rule is based on Green's theorem, which relates the double integral over the polygon to a line integral along the polygon boundary. Its efficiency can be greatly improved in our specific case by taking analytical advantage of the isotropy of f_2, after which numerical integration remains in only one dimension [see Meyer and Held (2014), Section 2.4]. Regardless of any sophisticated cubature rule, the required integration of f_2 over n polygons in the log-likelihood is the part that makes model fitting cumbersome: it introduces numerical errors which have to be controlled such that they do not corrupt numerical likelihood maximization, and it increases computational cost by several orders of magnitude. For instance, in our IMD application in Section 4.1 a single likelihood evaluation would only take 0.02 seconds if we used a constant spatial interaction function $f(x) \equiv 1$, where the integral does not depend on parameters being optimised and simply equals the area of the polygonal domain. For the Gaussian kernel, a single evaluation takes about 5 seconds, the step function takes 7 seconds, and the power law and Student kernel take about 20 seconds. The above and all following runtime statements refer to total CPU time at 2.80GHz (real elapsed time is shorter since some computations run in parallel on multiple CPUs).

3. Count data model.

3.1. *Introduction*. The multivariate time-series model established by Held and Paul (2012) [see also Held, Höhle and Hofmann (2005), Paul and Held (2011), Paul, Held and Toschke (2008)] is designed for spatially and temporally aggregated surveillance data, that is, disease counts Y_{it} in regions $i = 1, \ldots, I$ and periods $t = 1, \ldots, T$. Formally, the counts Y_{it} are assumed to follow a negative binomial distribution

$$Y_{it} | \mathbf{Y}_{\cdot,t-1} \sim \text{NegBin}(\mu_{it}, \psi), \qquad i = 1, \ldots, I, t = 1, \ldots, T$$

with additively decomposed mean

$$\mu_{it} = \nu_{it} e_{it} + \lambda_{it} Y_{i,t-1} + \phi_{it} \sum_{j \neq i} w_{ji} Y_{j,t-1}, \tag{7}$$

and overdispersion parameter ψ such that the conditional variance of Y_{it} is $\mu_{it}(1+\psi\mu_{it})$. The Poisson distribution results as a special case if $\psi = 0$. In (7), the first term represents the endemic component similar to the point process model (1). The endemic mean is proportional to an offset of known expected counts e_{it} typically reflecting the population at risk. The other two components are observation-driven epidemic components: an autoregression on the number of cases at the previous time point, and a "spatio-temporal" component capturing transmission from other units. Note that without these epidemic components, the model would reduce to a negative binomial regression model for independent observations.

Each of ν_{it}, λ_{it}, and ϕ_{it} is a log-linear predictor of the form

$$\log(\cdot_{it}) = \alpha^{(\cdot)} + b_i^{(\cdot)} + \boldsymbol{\beta}^{(\cdot)\top}\mathbf{z}_{it}^{(\cdot)}$$

(where "$\cdot$" is one of ν, λ, ϕ), containing fixed and region-specific intercepts as well as effects of exogenous covariates $\mathbf{z}_{it}^{(\cdot)}$ including time effects. For example, in the influenza application in Section 4.2,

$$\mathbf{z}_{it}^{(\nu)} = \big(t, \sin(1\cdot\omega t), \cos(1\cdot\omega t), \ldots, \sin(S\cdot\omega t), \cos(S\cdot\omega t)\big)^\top$$

describes an endemic time trend with a superposition of S harmonic waves of fundamental frequency $\omega = 2\pi/52$ [Held and Paul (2012)]. The random effects $\mathbf{b}_i := (b_i^{(\lambda)}, b_i^{(\phi)}, b_i^{(\nu)})^\top$ account for heterogeneity between regions, and are assumed to follow independently a trivariate normal distribution with mean zero and covariance matrix $\boldsymbol{\Sigma}$. Accounting for correlation of random effects across regions is possible by adopting a conditional autoregressive (CAR) model [Paul and Held (2011)].

The weights w_{ji} of the spatio-temporal component in (7) describe the strength of transmission from region j to region i, collected into an $I \times I$ weight matrix (w_{ji}). In contrast to the individual-level model, all of the $Y_{j,t-1}$ cases of the neighbour j by aggregation contribute with the same weight w_{ji} to infections in region i. In previous work, these weights were assumed to be known and restricted to first-order neighbours:

$$w_{ji} = \begin{cases} 1/n_j, & \text{for } i \sim j, \\ 0, & \text{otherwise,} \end{cases} \tag{8}$$

where the symbol "$\sim$" denotes "is adjacent to" and n_j is the number of direct (first-order) neighbours of region j. This is a normalised version of the "raw" adjacency indicator matrix $\mathbf{A} = (\mathbb{1}(i \sim j))_{j,i=1,\ldots,I}$, which is binary and symmetric. The idea behind normalisation is that each region j distributes its cases uniformly to its n_j neighbours [Paul, Held and Toschke (2008)]. Accordingly, the weight matrix is normalised to proportions such that all rows sum to 1. A simple alternative weight matrix considering only first-order neighbours would result from the definition $w_{ji} = 1/n_i$ for $i \sim j$ (i.e., columns sum to 1), meaning that the number of cases in a region i at time t is promoted by the mean of the neighbours at time $t-1$.

However, the first definition seems more natural in the framework of branching processes, where the point of view is from the infective source. Furthermore, the factor $1/n_i$ would be confounded with the region-specific effects $b_i^{(\phi)}$.

In either case, with the above weight matrix, the epidemic can only spread to first-order neighbours during the period $t \to t+1$, except for independently imported cases via the endemic component. This ignores the ability of humans to travel further. In what follows, we propose a parametric generalisation of the neighbourhood weights: a power law.

3.2. *Power-law extension.* To implement the power-law principle in the network of geographical regions, we first need to define a distance measure on which the power law acts. There are two natural choices: Euclidean distance between centroid coordinates and the order of neighbourhood. The first one conforms to a continuous power law, whereas the second one is discrete. However, using centroid coordinates interferes with the area and shape of the regions. Specifically, a tiny neighbouring region would be attributed a stronger link than a large neighbour with centroid further apart, even if the latter shares more boundary than the tiny region. Using the common boundary length as a measure of "coupling" [Keeling and Rohani (2002)] would only cover adjacent regions. We thus opt for the discrete measure of neighbourhood order.

Formally, a region j is a kth-order neighbour of another region i, denoted $o_{ji} = o_{ij} = k$, if it is adjacent to a $(k-1)$th-order neighbour of i and if it is not itself a neighbour of order $k-1$ of region i. In other words, two regions are kth-order neighbours, if the shortest route between them has k steps across distinct regions. The network of regions thus features a symmetric $I \times I$ matrix of neighbourhood orders with zeroes on the diagonal by convention.

Given this discrete distance measure, we generalise the previously used first-order weight matrix to higher-order neighbours assuming a power law with decay parameter $d > 0$:

$$w_{ji} = o_{ji}^{-d} \tag{9}$$

for $j \neq i$ and $w_{jj} = 0$. This may also be recognised as the kernel of the Zipf (1949) probability distribution. The raw power-law weights (9) can be normalised to

$$w_{ji} = \frac{o_{ji}^{-d}}{\sum_{k=1}^{I} o_{jk}^{-d}} \tag{10}$$

such that $\sum_{k=1}^{I} w_{jk} = 1$ for all rows j of the weight matrix. The higher the decay parameter d, the less important are higher-order neighbours. The limit $d \to \infty$ corresponds to the previously used first-order dependency, whereas $d = 0$ would assign equal weight to all regions.

Similarly to the point process modelling in Section 2.2, we also estimate the weights in a qualitative way by treating the order of neighbourhood as a factor:

$$w_{ji} = \sum_{o=1}^{M-1} \omega_o \cdot \mathbb{1}(o_{ji} = o) + \omega_M \cdot \mathbb{1}(o_{ji} \geq M). \tag{11}$$

Aggregation of higher orders ($o_{ji} \geq M$) is necessary since the available information becomes increasingly sparse. As before, the unconstrained weights (11) can be normalised to $w_{ji} / \sum_{k=1}^{I} w_{jk}$.

3.3. *Inference*. We set $\omega_1 = 1$ for identifiability and estimate the decay parameter d and the unconstrained weights $\omega_2, \ldots, \omega_M$ on the log-scale to enforce positivity. Supplied with the enhanced score function and Fisher information matrix, estimation of parametric weights is still possible within the penalised likelihood framework established by Paul and Held (2011) [see also Meyer and Held (2014), Section 1.2]. The authors argue, however, that classical model choice criteria such as Akaike's Information Criterion (AIC) cannot be used straightforwardly for models with random effects. Therefore, performance of the power-law models and the previous first-order formulations is compared by one-step-ahead forecasts assessed with strictly proper scoring rules: the logarithmic score (logS) and the ranked probability score (RPS) advocated by Czado, Gneiting and Held (2009) for count data:

$$\begin{aligned} \text{logS}(P, y) &= -\log P(Y = y), \\ \text{RPS}(P, y) &= \sum_{k=0}^{\infty} \big[P(Y \leq k) - \mathbb{1}(y \leq k)\big]^2. \end{aligned}$$

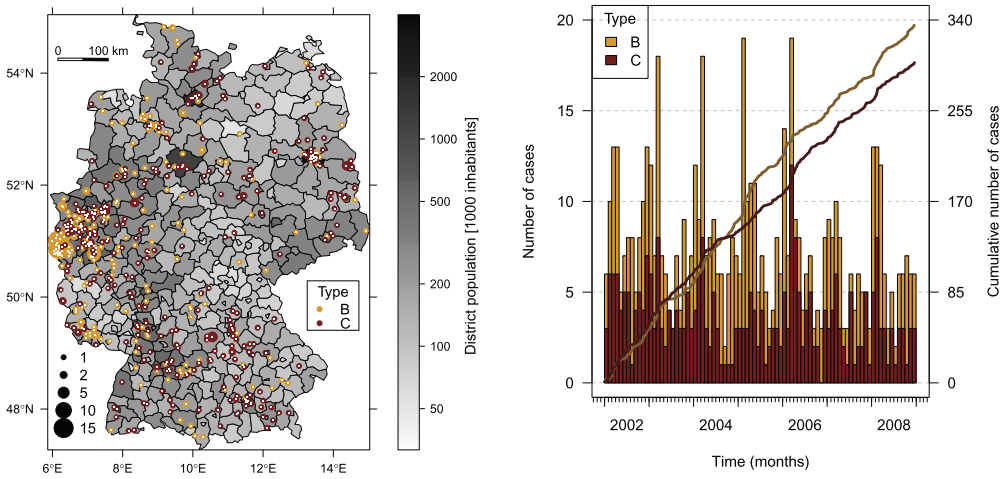
These scores evaluate the discrepancy between the predictive distribution P from a fitted model and the later observed value y. Thus, lower scores correspond to better predictions. Note that the infinite sum in the RPS can be approximated by truncation at some large k in a way such that a prespecified absolute approximation error is maintained [Wei and Held (2014)]. Such scoring rules have already been used for previous analyses of the influenza surveillance data [Held and Paul (2012)]. Along these lines, one-step-ahead predictions and associated scores are computed and statistical significance of the difference in mean scores is assessed using a Monte-Carlo permutation test for paired data.

4. Applications. We now apply the power-law formulations of both model frameworks to previously analysed surveillance data and investigate potential improvements with respect to predictive performance. We investigate the appropriateness of the power-law shape by alternative qualitative estimates of spatial interaction. In Section 4.1 635 individual case reports of IMD caused by the two most common bacterial finetypes of meningococci in Germany from 2002 to 2008

are analysed with the point process model (1). In Section 4.2 the multivariate time-series model (7) is applied to weekly numbers of reported cases of influenza in the 140 administrative districts of the federal states Bavaria and Baden-Württemberg in Southern Germany from 2001 to 2008. In Section 4.3 we evaluate a simulation-based long-term forecast of the 2008 influenza wave. Space–time animations of both surveillance data sets are provided in Supplement A.

4.1. *Cases of invasive meningococcal disease in Germany*, 2002–2008 (*see Figure 2*). In the original analysis of the IMD data [Meyer, Elias and Höhle (2012)], comprehensive AIC-based model selection yielded a linear time trend, a sinusoidal time-of-year effect ($S = 1$), and no effect of the (lagged) number of local influenza cases in the endemic component. The epidemic component included an effect of the meningococcal finetype (C:P1.5,2:F3-3 being less infectious than B:P1.7-2,4:F1-5, abbreviated by C and B in the following), a small age effect (3–18 year old patients tending to be more infectious), and supported an isotropic Gaussian spatial interaction function f compared to a homogeneous spatial spread [$f(x) \equiv 1$]. The analysis assumed constant infectivity over time until 30 days after infection when infectivity vanishes to zero, that is, $g(t) = \mathbb{1}_{(0,30]}(t)$. In this paper, we replace the Gaussian kernel in the selected model by the proposed power-law distance decay (3) to investigate if it better captures the dynamics of IMD spread.

Note that the distinction between two finetypes in this application actually corresponds to a marked version of the point process model. It is described by an in-



(a) Spatial point pattern with dot size proportional to the number of cases at the respective location (postcode level)

(b) Monthly aggregated time series and evolution of the cumulative number of cases (by date of specimen sampling)

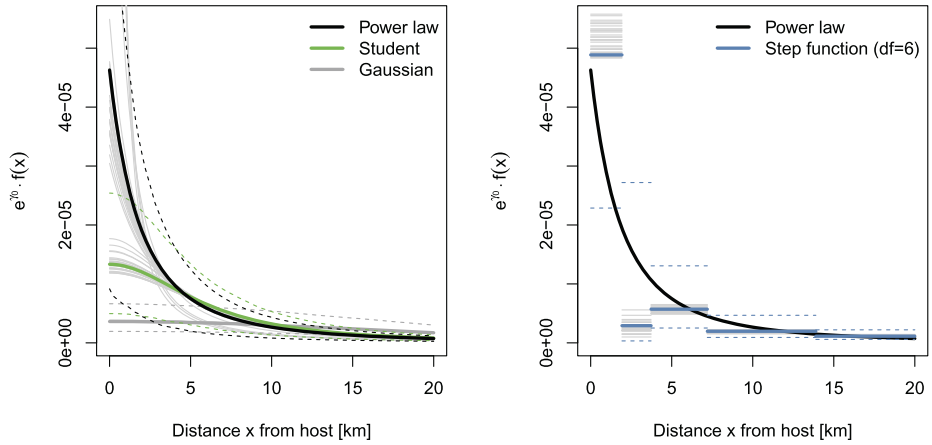
FIG. 2. *Distribution of the* 635 *IMD cases in Germany*, 2002–2008, *caused by the two most common meningococcal finetypes B:P*1.7-2,4:*F*1-5 (335 *cases*) *and C:P*1.5,2:*F*3-3 (300 *cases*), *as reported to and typed by the German Reference Centre for Meningococci.*

tensity function $\lambda(t, \mathbf{s}, k)$, where the sum in (1) is restricted to previously infected individuals with bacterial finetype k, since we assume that infections of different finetypes are not associated via transmission [Meyer, Elias and Höhle (2012)]. For convenience, we kept notation simple and comparable to the multivariate time-series model of Section 3.

Prior to fitting point process models to the IMD data, the interval-censored nature of the data caused by a restricted resolution in space and time has to be taken into account: we only observed dates and residence postcodes of the cases (implicitly assuming that infections effectively happened within the residential neighbourhood). This makes the data interval-censored, yielding tied observations. However, ties are not compatible with our (continuous-time, continuous-space) point process model since observing two events at the exact same time point or location has zero probability. In the original analyses with a Gaussian kernel f, events were untied in time by subtracting a $U(0, 1)$-distributed random number from all observed time points [Meyer, Elias and Höhle (2012)], that is, random sampling within each day, which is also the preferred method used by Diggle, Kaimi and Abellana (2010). To identify the two-parameter power law $(x + \sigma)^{-d}$, it was additionally necessary to break ties in space, since otherwise $\log \sigma$ diverged to $-\infty$, yielding a pole at $x = 0$. A possible solution is to shift all locations randomly in space within their round-off intervals similar to the tie-breaking in time. Lacking a shapefile of the postcode regions, we shifted locations by a vector uniformly drawn from the disc with radius $\varepsilon/2$, where ε is the minimum observed spatial separation of distinct points, here $\varepsilon = 1.17$ km. Accordingly, a sensitivity analysis was conducted by applying the random tie-breaking in time and space 30 times and fitting the models to all replicates.

Figure 3(a) displays estimated spatial interaction functions—appropriately scaled by $\exp(\hat{\gamma}_0)$—together with confidence intervals and estimates from the sensitivity analysis (see Table 1 for values of $\hat{\gamma}_0$, $\hat{\sigma}$, and $\hat{d}$). The power law puts much more weight on localised transmissions with an initially faster distance decay of infectivity. Furthermore, it features a heavier tail than the Gaussian kernel, which facilitates the geographical spread of IMD by occasional long-range transmissions. Maps of the accumulated epidemic intensity [Meyer and Held (2014), Figure 1] visualise the impact of the power law on the modelled infectivity. Sensitivity analysis shows that AIC clearly prefers the new power-law kernel against the Gaussian kernel (mean Δ AIC $= -27.6$, SD $= 1.5$). The Student kernel represents a compromise between the other two parametric kernels with short-range properties similar to the Gaussian kernel but with a heavy tail. However, AIC improvement is not as large as for the above power law (mean Δ AIC $= -15.5$, SD $= 0.9$).

For these three kernels, sensitivity analysis of the random tie-breaking procedure in space and time generally confirmed the results. The Gaussian kernel was least affected by the small-scale perturbation of event times and locations. Some replicates for the power-law model yielded a slightly steeper shape, which is due to closely located points after random tie-breaking. Such an artifact would have been



(a) Power laws vs. the Gaussian kernel (b) Power law vs. a step function

FIG. 3. *Estimated spatial interaction functions—appropriately scaled by the epidemic intercept* $\exp(\gamma_0)$. *The dashed lines represent* 95% *confidence intervals obtained as the pointwise* 2.5% *and* 97.5% *quantiles of the functions evaluated for* 999 *samples from the asymptotic multivariate normal distribution of the affected parameters. The light grey lines are estimates obtained from a sensitivity analysis with repeated random tie-breaking.*

avoided if we had used constrained sampling in that the randomly shifted points obey a minimum separation of say 0.1 km.

The estimated lagged version of the power law (4) is shown in Supplement B [Meyer and Held (2014), Figure 2]. It has a uniform short-range dispersal radius

TABLE 1

Parameter estimates and 95% *Wald confidence intervals for the Gaussian and the power-law model. Results for the Gaussian kernel are slightly different from those reported by Meyer, Elias and Höhle (2012) due to improved numerical integration. Note that we use the symbol* σ *for the scale parameter and* d *for the decay parameter in all spatial interaction functions, but these parameters as well as* γ_0 *are not directly comparable (instead see Figure 3)*

	Gaussian kernel (2)		**Power-law kernel (3)**	
	Estimate	**95% CI**	**Estimate**	**95% CI**
β_0	−20.53	−20.62 to −20.44	−20.58	−20.68 to −20.47
β_{trend}	−0.05	−0.09 to −0.00	−0.05	−0.09 to 0.00
β_{sin}	0.26	0.14 to 0.39	0.26	0.12 to 0.39
β_{cos}	0.26	0.14 to 0.39	0.27	0.14 to 0.40
γ_0	−12.53	−13.15 to −11.91	−6.21	−9.32 to −3.10
γ_{C}	−0.91	−1.44 to −0.39	−0.80	−1.31 to −0.29
γ_{3-18}	0.67	0.04 to 1.31	0.78	0.11 to 1.45
$\gamma_{\geq 19}$	−0.29	−1.19 to 0.61	−0.18	−1.11 to 0.75
σ	16.37	13.95 to 19.21	4.60	1.80 to 11.71
d			2.47	1.80 to 3.39

of $\hat{\sigma} = 0.40$ (95% CI: 0.18 to 0.86) kilometres. However, such a small σ is not interpretable since it is actually not covered by the spatial resolution of the data. Accordingly, the 30 estimates of the sensitivity analysis are more dispersed, as is the goodness of fit compared to the Gaussian kernel (mean $\Delta\,\mathrm{AIC} = -21.1$, $\mathrm{SD} = 3.8$).

Figure 3(b) shows a comparison of the estimated power law with a step function (6) for spatial interaction. An upper boundary knot had to be specified, which we set at 100 kilometres, where the step function drops to 0. We chose six knots to be equidistant on the log-scale within $[0, \log(100)]$, that is, steps at 1.9, 3.7, 7.2, 13.9, 26.8, and 51.8 kilometres. Estimation took only 72 seconds due to the analytical implementation of the integration of f_2 over polygonal domains, whereas the power-law model took 42 minutes. The power law is well confirmed by the step function; it is almost completely enclosed by its 95% confidence interval. The step function suggests an even steeper initial decay and has a slightly better fit in terms of AIC (mean $\Delta\,\mathrm{AIC} = -6.9$, $\mathrm{SD} = 4.0$ compared to the power law). However, it depends on the choice of knots, it is sensitive for artifacts of the data and forfeits monotonicity, which contradicts Tobler's *first law of geography* [Tobler (1970)].

Parameter estimates and confidence intervals for the Gaussian and the power-law model are presented in Table 1 [see Meyer and Held (2014), Table 1, for parameter estimates of the other models]. The parameters of the endemic component characterising time trend and seasonality were not affected by the change of the shape of spatial interaction, and also the epidemic coefficients of finetype and age group do not differ much between the models retaining their signs and orders of magnitude. For instance, also with the power-law kernel, the C-type is approximately half as infectious as the B-type, which is estimated by the multiplicative type-effect $\exp(\hat{\gamma}_{\mathrm{C}}) = 0.45$ (95% CI: 0.27 to 0.75) on the force of infection (type B is the reference category here).

An important quantity in epidemic modelling is the expected number R of offspring (secondary infections) each case generates. This reproduction number can be derived from the fitted models for each event by integrating its triggering function $\eta_j g(t - t_j) f(\|\mathbf{s} - \mathbf{s}_j\|)$ over the observation region $\mathbf{W}$ and period $[t_j, T]$ [Meyer, Elias and Höhle (2012)]. Type-specific estimates of R are then obtained by averaging over the individual estimates by finetype. Table 2 shows that the reproduction numbers become slightly larger, which is related to the heavier tail of the power law enabling additional interaction between events at far distances.

We close this application with two additional ideas for improvement of the model. First, it might be worth considering a population effect also in the *epidemic* component to reflect higher contact rates and thus infectivity in regions with a denser population. Using the log-population density of the infective's district, $\log \rho_{[t_j][\mathbf{s}_j]}$, the corresponding parameter is estimated to be $\hat{\gamma}_{\log(\rho)} = 0.21$ (95% CI: -0.07 to 0.48), that is, individual infectivity scales with $\rho^{0.21}$, where ρ ranges from 39 to 4225 km^2. Although the positive point estimate supports this

TABLE 2
Type-specific reproduction numbers with 95% *confidence intervals* (*based on* 199 *samples from the asymptotic multivariate normal distribution of the parameter estimates*)

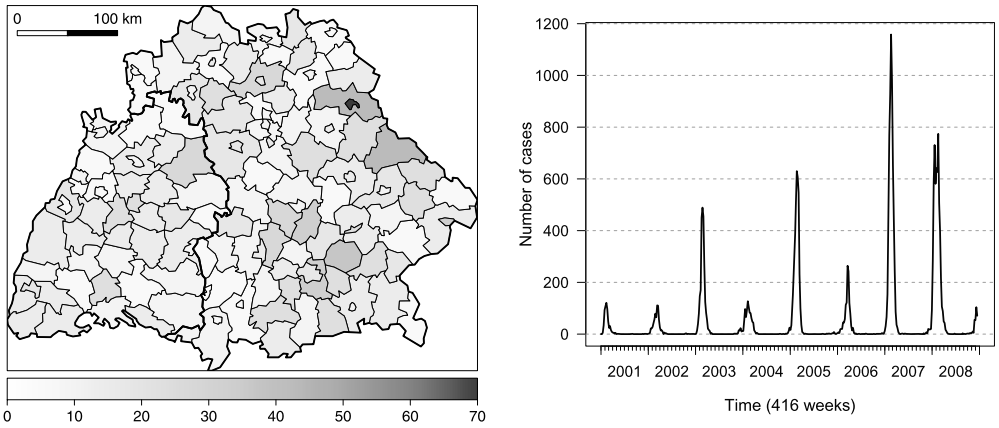
	Gaussian kernel (2)		**Power-law kernel (3)**	
	Estimate	**95% CI**	**Estimate**	**95% CI**
B	0.22	0.17 to 0.31	0.26	0.10 to 0.35
C	0.10	0.06 to 0.15	0.13	0.05 to 0.19

idea, the wide confidence interval does not reflect strong evidence for such a population effect in the IMD data.

However, it is helpful to allow for spatial heterogeneity in the *endemic* component. For instance, an indicator for districts at the border or the distance of the district's centroids from the border could serve as proxies for simple edge effects. The idea is that as we get closer to the edge of the observation window (Germany) more infections will originate from external sources not directly linked to the observed history of the epidemic within Germany. We thus model a spatially varying risk of importing cases through the endemic component. For the Greater Aachen Region in the central-west part of Germany, where a spatial disease cluster is apparent in Figure 2(a), such a cross-border effect with the Netherlands was indeed identified by Elias et al. (2010) for the serogroup B finetype during our observation period using molecular sequence typing of bacterial strains in infected patients from both countries. Inclusion of an edge indicator in the endemic covariates $\mathbf{z}_{[t][\mathbf{s}]}$ improves AIC by 5 with an estimated rate ratio of 1.37 (95% CI: 1.10 to 1.70) for districts at the border versus inner districts. If we instead use the distance to the border, AIC improves by 20 with an estimated risk reduction of 5.0% (95% CI: 3.0% to 7.0%) per 10 km increase in distance to the border.

4.2. *Influenza surveillance data from Southern Germany*, 2001–2008 (*see Figure* 4). The best model (with respect to logS and RPS) for the influenza surveillance data found by Held and Paul (2012) using normalised first-order weights included $S = 1$ sinusoidal wave in each of the autoregressive (λ_{it}) and spatio-temporal (ϕ_{it}) components and $S = 3$ harmonic waves with a linear trend in the endemic component ν_{it} with the population fraction e_i in region i as offset. We now fit an extended model by estimating (raw or normalised) power-law neighbourhood weights (9) or (10) as described in Section 3.2, which replace the previously used fixed adjacency indicator.

Figure 5(a) shows the estimated normalised power law with $\hat{d} = 1.80$ (95% CI: 1.61 to 2.01). This decay is remarkably close to the power-law exponent 1.59 estimated by Brockmann, Hufnagel and Geisel (2006) for short-time travel in the USA with respect to distance (in kilometres), even though neighbourhood order is a discretised measure with no one-to-one correspondence to Euclidean distances, and

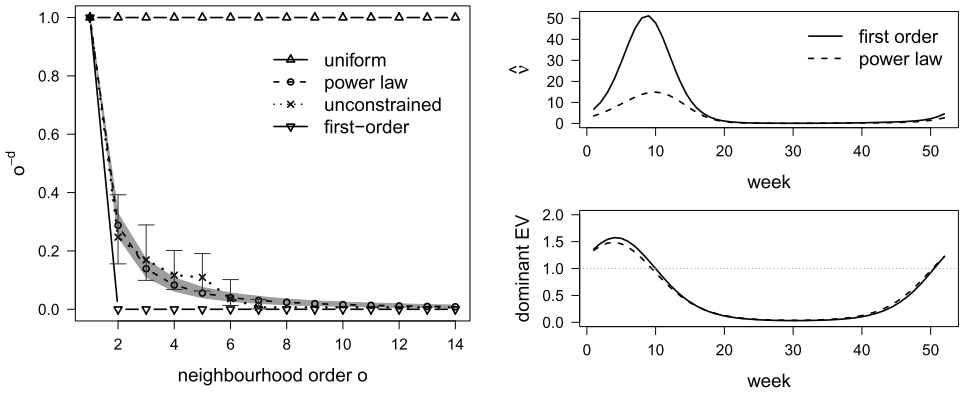


(a) Mean yearly incidence per 100,000 inhabitants

(b) Weekly number of cases

FIG. 4. *Spatial and temporal distribution of reported influenza cases in the* 140 *districts of Bavaria and Baden-Württemberg during the years* 2001 *to* 2008.

travel behaviour in the USA is potentially different from that in Southern Germany. The plot also shows the estimated unconstrained weights for comparison with the power law. The sixth order of neighbourhood was the highest for which we could estimate an individual weight; higher orders had to be aggregated corresponding to $M = 7$ in (11). The unconstrained weights decrease monotonically and resemble nicely the estimated power law, which is enclosed by the 95% confidence intervals (except for order 5, which has a slightly higher weight). The results with raw weights are very similar and shown in Meyer and Held [(2014), Figure 3].



(a) Power-law (10) and unconstrained weights (11) with 95% confidence intervals

(b) Seasonal variation of the endemic (top) and epidemic (bottom) components

FIG. 5. *Estimated power-law and unconstrained weights* (a), *and seasonal variation* (b) *using normalised weights*.

Figure 5(b) shows the estimated seasonal variation in the endemic component and the course of the dominant eigenvalue [Held and Paul (2012)] for the normalised weight models. The dominant eigenvalue is a combination of the two epidemic components: if it is smaller than 1, it can be interpreted as the epidemic proportion of total disease incidence, otherwise it indicates an outbreak period. Whereas the course of this combined measure is more or less unchanged upon accounting for higher-order neighbours with a power law, the weight of the endemic component decreases remarkably. This goes hand in hand with an increased importance of the spatio-temporal component since in the power-law formulation much more information can be borrowed from the number of cases in other regions. Jumps of the epidemic to nonadjacent regions within one week are no longer dedicated to the endemic component only.

Concerning the remaining coefficients, there is less overdispersion in the power-law models (see ψ in Table 3), which indicates reduced residual heterogeneity. For the variance and correlation estimates of the random effects, there is no substantial difference between first-order and power-law models and even less between raw and normalised formulations.

To assess if the power-law formulation improves the previous first-order model, their predictive performance is compared based on one-week-ahead predictions for all 140 regions and the 104 weeks of the last two years. Computing these predictions for one model takes about 3 hours, since it needs to be refitted for every time point. Table 4 shows the resulting mean scores with associated p-values. Both

TABLE 3

Estimated model parameters (with standard errors) excluding intercepts and trend/seasonal coefficients. The parameter $\beta^{(\phi)}_{\log(\text{pop})}$ in the first row belongs to a further extended power-law (PL) model, which accounts for population in the spatio-temporal component (last column). The $\sigma^2_{\cdot}$ and $\rho_{\cdot\cdot}$ parameters are the variances and correlations of the random effects (from $\boldsymbol{\Sigma}$). The last row shows the final values of the penalised and marginal log-likelihoods

	Raw weights		**Normalised weights**		
	First order	**Power law**	**First order**	**Power law**	**PL + pop.**
$\beta^{(\phi)}_{\log(\text{pop})}$	–	–	–	–	0.76 (0.13)
d	–	1.72 (0.10)	–	1.80 (0.10)	1.65 (0.10)
ψ	0.93 (0.03)	0.86 (0.03)	0.92 (0.03)	0.86 (0.03)	0.86 (0.03)
σ^2_λ	0.14	0.17	0.13	0.17	0.16
σ^2_ϕ	0.94	0.92	0.98	0.89	0.71
σ^2_ν	0.50	0.67	0.51	0.67	0.66
$\rho_{\lambda\phi}$	0.02	0.20	0.03	0.21	0.13
$\rho_{\lambda\nu}$	0.11	0.31	0.12	0.31	0.27
$\rho_{\phi\nu}$	0.56	0.29	0.55	0.30	0.39
$l_{\text{pen}}(l_{\text{mar}})$	−18,400 (−433)	−18,129 (−456)	−18,387 (−436)	−18,124 (−453)	−18,124 (−439)

TABLE 4

Mean scores of 104×140 *one-week-ahead predictions over the last two years, accompanied with p-values for comparing power-law and first-order weights obtained via permutation tests with* 19,999 *random permutations. Note that the values obtained for normalised first-order weights are slightly different from the ones published by Held and Paul (2012) due to a correction of a recording error in the last week of the influenza data*

	Raw weights		**Normalised weights**	
	logS	**RPS**	**logS**	**RPS**
First order	0.5522	0.4205	0.5511	0.4194
Power law	0.5453	0.4174	0.5448	0.4168
p-value	0.00005	0.11	0.0001	0.19

logS and RPS improve when accounting for higher-order neighbours with a power law, while the difference is only significant for the logarithmic score. Furthermore, the normalised formulation performs slightly better than the raw weights. For instance, the mean difference in the logarithmic scores of the respective power-law models has an associated p-value of 0.0009. In the following we therefore only consider the normalised versions. For additional comparison, the simple uniform weight model ($w_{ji} \equiv 1$), which takes into account higher-order neighbours but with equal weight, has mean logS $= 0.5484$ and mean RPS $= 0.4215$, and thus performs worse than a power-law decay and, according to the RPS, even worse than first-order weights.

Similarly to the IMD analysis, further improvement of the model's description of human mobility can be achieved by accounting for the district-specific population also in the spatio-temporal component. The idea is that there tends to be more traffic to regional conurbations, that is, districts with a larger population, which are thus expected to import a bigger amount of cases from neighbouring regions [Bartlett (1957)]. Note that inclusion of the log-population in $\mathbf{z}_{it}^{(\phi)}$ affects susceptibility rather than infectivity, which is inverse to modelling the force of infection in the individual-based framework. The influenza data yield an estimated coefficient of $\hat{\beta}_{\log(\text{pop})}^{(\phi)} = 0.76$ (95% CI: 0.50 to 1.01), which provides strong evidence for such an agglomeration effect. The variance of the random effect $b_i^{(\phi)}$ of the spatio-temporal component is slightly reduced from 0.89 to 0.71, reflecting a decrease in residual heterogeneity between districts. The decay parameter is estimated to be slightly smaller in the extended model [$\hat{d} = 1.65$ (95% CI: 1.45 to 1.86)] and all other effects remain approximately unchanged (see Table 3). However, the predictive performance improves only minimally, for example, the logarithmic score decreases from 0.5448 to 0.5447 ($p = 0.66$). This small change could be related to the random effects $b_i^{(\phi)}$, which replace parts of the population effect if it is not included as a covariate. Indeed, there is correlation ($r_{\text{Pearson}} = 0.41$) between

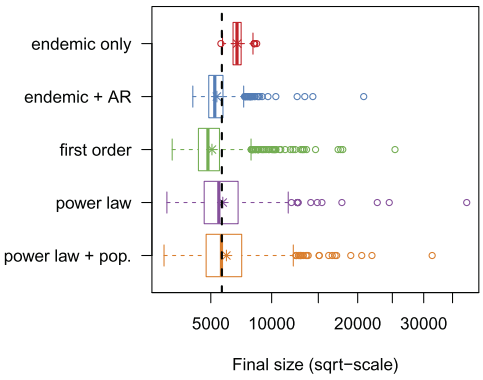
$\log(\text{pop}_i)$ and $b_i^{(\phi)}$ in the model without an explicit population effect in ϕ_{it} [see the scatterplot in Meyer and Held (2014), Figure 5].

4.3. *Long-term forecast of the* 2008 *influenza wave.* For further evaluation of the power-law models described in Section 4.2, we carry out a long-term forecast of the wave of influenza in 2008. Specifically, we simulate the evolution of the epidemic during the first 20 weeks in 2008 for each model trained by the previous years and initialised by the 18 cases of the last week of 2007 (see the animation in Supplement A, for their spatial distribution). Predictive performance is then evaluated by the final size distributions and by proper scoring rules assessing the empirical distributions induced by the simulated counts both in the temporal and spatial domains. Since the logarithmic score is infinite in the case of zero predictive probability for the observed count, we instead use the Dawid and Sebastiani (1999) score

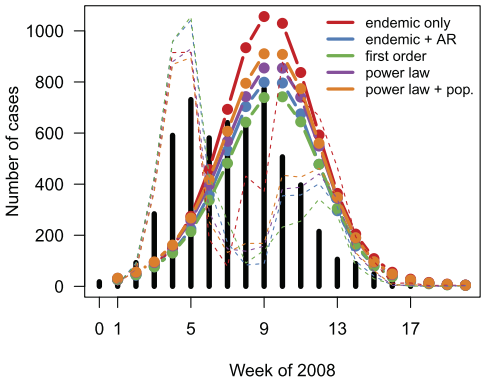
$$\text{DSS}(P, y) = \frac{(y - \mu_P)^2}{\sigma_P^2} + \log \sigma_P^2,$$

where μ_P and σ_P^2 denote the mean and the variance of P [see also Gneiting and Raftery (2007)].

Figure 6(a) shows the final size distributions of the simulated waves of influenza during the first 20 weeks of 2008. Note that model complexity increases from top to bottom and that we also considered the naive endemic model, that is, independent counts, and the model without a spatio-temporal component as additional benchmarks. The endemic-only model, which decomposes disease incidence into spatial



(a) Final size distributions ($\sqrt{}$-scale). The star in each box represents the mean, and the vertical dashed line marks the observed final size of 5781 cases



(b) Time series of observed (bars) and mean simulated (dots) counts aggregated over all districts. Week 0 corresponds to the initial condition (2007-W52). The dashed lines show the (scaled) RPS (see also Table 5)

FIG. 6. *Summary statistics of* 1000 *simulations of the wave of influenza during the first* 20 *weeks of* 2008 *for five competing models.*

TABLE 5
Long-term predictive performance of 5 *competing models in the temporal and spatial dimensions measured by mean DSS and RPS for the* 2008 *wave of influenza*

Model	Time		Space		Space–time	
	DSS	RPS	DSS	RPS	DSS	RPS
Endemic only	27.03	149.77	7.85	15.39	2.91	1.31
Endemic + autoregressive	31.36	112.15	7.59	15.04	2.58	1.26
First order	26.46	108.61	7.51	15.63	2.50	1.26
Power law	16.41	110.20	7.36	14.75	2.29	1.25
Power law + population	15.49	111.86	7.24	14.30	2.29	1.24

variation across districts, a seasonal and a log-linear time trend, overestimates the reported size of 5781 cases. It also does not allow for much variability in the size of the outbreak as opposed to the models with epidemic potential. The power-law models show the greatest amount of variation but best meet the reported final size: the power-law model without the population effect yields a simulated mean of 6022 (95% CI: 3126 to 10,808). The huge uncertainty seems plausible with regard to the long forecast horizon over a whole epidemic wave.

Figure 6(b) shows the time series of observed and mean simulated counts aggregated over all districts. In 2008, the wave grew two or more weeks earlier than in previous years trained by the sinusoidal terms in the three components. This phenomenon cannot be captured by the simulations, which are solely based on the observed pattern during 2001–2007 and the distribution of the cases from the last week of 2007. Furthermore, instead of two peaks as observed specifically in 2008, the simulations yield a single, larger peak where the power-law models on average induce the best amplitudes with respect to final size. The simulated spatial distribution of the cases (see Figure 7) is very similar among the various models and agrees quite well with the observed pattern. Animations of the observed and mean simulated epidemics provide more insight about the epidemic spread and are available in Supplement A. It is difficult to see a clear-cut traveling-wave of influenza in the reported data, which suggests that both an endemic component capturing immigration as well as scale-free jumps via the spatio-temporal component, that is, power-law weights w_{ji}, are important. Supplement A also includes an animated series of weekly probability integral transform (PIT) histograms [Gneiting, Balabdaoui and Raftery (2007)] using the nonrandomised version for count data proposed by Czado, Gneiting and Held (2009). These sequential PIT histograms mainly reflect the above time shift of the predictions. More clearly than the plots, the mean scores in Table 5 show that predictive performance generally improves with increasing model complexity and use of a power-law decay.

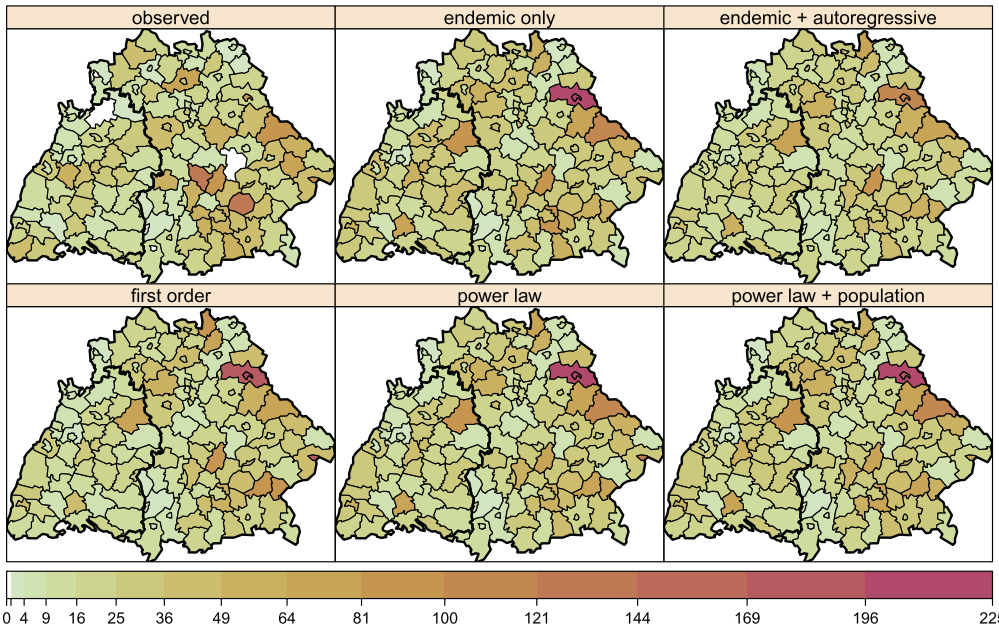


FIG. 7. *Observed and mean simulated incidence* (*cases per* 100 000 *inhabitants*) *aggregated over the* 20 *weeks forecast horizon* (*see Figure* 6 *of Supplement B for scatterplots*).

5. Discussion. Motivated by the finding of Brockmann, Hufnagel and Geisel (2006) that short-time human travel roughly follows a power law with respect to distance, we investigated a power-law decay of spatial dependence between infections in two modelling frameworks for spatio-temporal surveillance data. A spatio-temporal point process model was applied to case reports of invasive meningococcal disease, and a multivariate time-series model was applied to counts of influenza aggregated by week and district. Since human mobility is an important driver of epidemic spread, the aim was to improve the predictive performance of these models using a power-law transmission kernel with respect to distance or neighbourhood order, respectively, where the decay is estimated jointly with all other model parameters.

In both applications considered, the power-law formulations performed better than previously used naive Gaussian or first-order interaction models, respectively. Furthermore, alternative piecewise constant, but otherwise unrestricted interaction models were in line with the estimated power laws. This confirms that the power-law distribution of short-time human travel translates to the modelling of infectious disease spread. We note that the qualitative interaction models could be replaced by (cubic) smoothing spline formulations, either in a continuous [Eubank (2000)] or in discrete fashion [Fahrmeir and Knorr-Held (2000)]. In order to penalise deviations from the power law, this should be done on a log–log scale, where the power law is a simple linear relationship. However, data-driven estimation of the smoothing parameter may become difficult.

The heavy tail of the power law allows for long-range dependence between cases, which accordingly increased the importance of the epidemic component in both models. An alternative formulation of spatial interaction with occasional long-range transmission was used by Diggle (2006), who added a small distance-independent value to a powered exponential term of the scaled distance. However, this offset and the power parameter are poorly identified. For the 2001 UK foot-and-mouth disease epidemic, Keeling et al. (2001) observed a power-law-like, sharply peaked transmission kernel, and Chis Ster and Ferguson (2007) subsequently found that the power law (3) yields a much better fit than the offset kernel or other functional forms, which is in accordance with our results for the spread of human infectious diseases.

Regions at the edge of the observation window are missing potential sources of infection from the unobserved side of the border. To capture unobserved heterogeneity due to immigration/edge effects, the count data model includes region-specific random effects $b_i^{(\nu)}$ in the endemic component. However, there was no clear pattern in their estimates with respect to regions being close to the border or not [Meyer and Held (2014), Figure 4]. In contrast, the IMD data supported edge effects, specifically concerning the border to the Netherlands. The spatial occurrence of cases met our simplistic approach of including the distance to the border as a covariate in the endemic component. This ignores that immigration might be more important in large metropolitan areas attracting people from abroad regardless of the location within Germany. A better way of accounting for edge effects would thus be to explicitly incorporate immigration data. For instance, Geilhufe et al. (2014) used incoming road or air traffic from outside North Norway as a proxy for the risk of importing cases of influenza, which led to improved predictive performance while also accounting for population in the spatio-temporal component.

Scaling regional susceptibility by population size proved very informative also for influenza in Southern Germany: more populated regions seem to attract more infections from neighbours than smaller regions, which reflects commuter-type imports [see Viboud et al. (2006), and Keeling and Rohani (2008), Section 6.3.3.1]. An exception of such a population effect in the spatio-temporal component might be seasonal accumulations in low-populated touristic regions. In the point process model for the IMD cases, the effect of population density on infectivity was less evident, which might be related to the very limited size of the point pattern with less than 100 cases per year over all of Germany.

Another limitation of the IMD data set is tied locations of cases due to censoring at the postcode level. For the power law to be identifiable, we randomly sampled the locations from discs of radius 0.59 km around the centroid of the respective postcode area, and verified that our results are insensitive to the random seed. Note that choosing a larger radius of, for example, 3 km, leads to less pronounced weight towards zero distance but yields otherwise similar results, especially concerning the relative performance of the various interaction functions.

We considered power laws as a description of spatial dispersal of infectious diseases as motivated by human travelling behaviour. Concerning temporal dispersal, power laws are usually not an appropriate description of the evolution of infectivity over time. Infectious diseases typically feature a very limited period of infectivity after the incubation period, since an infected individual will receive treatment and typically restrict its interaction radius upon the appearance of symptoms. Due to the small number of cases in the IMD data, we could not estimate a parametric temporal interaction function $g(t)$ and simply assumed constant infectivity during 30 days as in Meyer, Elias and Höhle (2012). More generally, $g(t)$ could represent an increasing level of infectivity beginning from exposure, followed by a plateau and then decreasing and eventually vanishing infectivity [Lawson and Leimich (2000), Section 5.3]. In the multivariate time-series model, the counts were restricted to only explicitly depend on the previous week. This is reasonable if the generation time, the time consumed by an infective to cause a secondary case, is not larger than the aggregation time in the surveillance data. For human influenza, Cowling et al. (2009) report a mean generation time of 3.6 days (95% CI: 2.9 days to 4.3 days).

Long-term simulated forecast of the 2008 influenza wave confirmed that the power-law model yields better predictions. However, the model was not able to describe the onset in 2008, which was two weeks earlier than in the years 2001–2007. For this to work, it would be necessary to further enrich the model by external processes such as vaccination coverage [as in Herzog, Paul and Held (2011)] or climate conditions [Fuhrmann (2010), Willem et al. (2012)] entering as covariates in the endemic and/or epidemic components. An alternative approach has been used by Fanshawe et al. (2008), where seasonality parameters were allowed to change from year to year according to a random walk model. Implementation would then require Markov chain Monte Carlo or other more demanding techniques for inference. Despite the open issue of dynamic seasonality, the simulated final size and spatial distribution matched the reported epidemic quite well.

This success also suggests that under-reporting of influenza was roughly constant over time. For instance, the 4 districts which did not report any cases during the 2008 forecast period (SK Kempten, SK Memmingen, LK Kelheim, and SK Aschaffenburg) only reported 1, 0, 20, and 4 cases in total during 2001–2007. However, we can only model the effectively reported number of cases, which may be affected by time-varying attention drawn to influenza in the media. Syndromic surveillance systems aim to unify various routinely collected data sources, for example, web searches for outbreak detection and monitoring [Hulth, Rydevik and Linde (2009), Josseran et al. (2006)], and may thereby provide a more realistic picture of influenza.

Prospective detection of outbreaks is also possible based on the count data model presented here. A statistic could be based on quantiles of the distribution of $Y_{i,t+1}|\mathbf{Y}_{\cdot,t}$, for example, an alarm could be triggered if the actual observed counts

at $t + 1$ are above the 99% quantile, say, [Held et al. (2006)]. Note that by including seasonality in the model, a yearly wave at the beginning of the year would be 'planned' and not necessarily considered a deviation from default behaviour.

Our power-law approach is very useful in the absence of movement network data (e.g., plane and train traffic). However, if such data were available [Lazer et al. (2009)], neighbourhood weights w_{ji} in the count data model could instead be based on the connectivity between regions, which was investigated by Schrödle, Held and Rue (2012) for the spread of Coxiellosis in Swiss cows and by Geilhufe et al. (2014) for the spread of influenza in Northern Norway. In recent work, Brockmann and Helbing (2013) introduce the 'effective distance' to describe the 2009 H1N1 influenza pandemic. Their approach relates to what has already been termed 'functional distance' by Brown and Horton (1970), that is, a function of (inter-)regional properties like population and commuter or travel flows such that it "reflects the net effect of entity properties upon the propensity of the entities to interact" [Brown and Holmes (1971)]. A recent example of using telephone call data as a measure of human interaction can be found in Ratti et al. (2010). Another fruitful area of future research is the statistical analysis of age-stratified surveillance data. Contact patterns vary across age [Mossong et al. (2008), Truscott et al. (2012)], calling for a unified analysis across age groups and regions.

APPENDIX: SOFTWARE

All calculations have been carried out in the statistical software environment R 3.0.2 [R Core Team (2013)]. Both model frameworks and their power-law extensions presented in this paper are implemented in the R package *surveillance* [Höhle, Meyer and Paul (2014)] as of version 1.6-0 available from the Comprehensive R Archive Network (CRAN.R-project.org). The two analysed data sets are included therein as `data("imdepi")` (courtesy of the German Reference Centre for Meningococci) and `data("fluBYBW")` [raw data obtained from the German national surveillance system operated by the Robert Koch Institute (2009)]. The point process model (1) for individual point-referenced data can be fitted by the function `twinstim()`, and the multivariate time-series model (7) for count data is estimated by `hhh4()`. The implementations are flexible enough to allow for other specifications of the spatial interaction function f and the weights w_{ji}, respectively. A related two-component epidemic model [Höhle (2009)], which is designed for time-continuous individual surveillance data of a closed population with a fixed set of locations, for example, for farm- or household-based epidemics, is also included as function `twinSIR()`. The application of all three model frameworks in R is described in detail in Meyer, Held and Höhle (2014).

Spatial integrals in the point process likelihood have been evaluated using cubature methods implemented in the R package *polyCub* 0.4-3 [Meyer (2014)]. Maps have been produced using *sp* 1.0-15 [Bivand, Pebesma and Gómez-Rubio (2013)] and animations using *animation* 2.2 [Xie (2013)].

Acknowledgements. This work was presented at the *Summer School on Topics in Space–Time Modeling and Inference* at Aalborg University, May 2013, which enabled fruitful discussions with its participants. These also gave rise to the efficient cubature rule for isotropic functions over polygonal domains elaborated in Meyer and Held [(2014), Section 2.4] with valuable support by Emil Hedevang and Christian Reiher. We thank Michaela Paul for technical support on the original count data model, as well as Johannes Elias and Ulrich Vogel from the German Reference Centre for Meningococci for providing us with the IMD data. We also appreciate helpful comments by Julia Meyer, Michael Höhle, the Editor Tilmann Gneiting, and two anonymous referees.

SUPPLEMENTARY MATERIAL

Supplement A: Animations of the IMD and influenza epidemics (http://www.biostat.uzh.ch/static/powerlaw/).

- Observed evolution of the IMD and influenza epidemics.
- Simulated counts from various models for the 2008 influenza wave.
- Weekly mean PIT histograms for these predictions.

Supplement B: Inference details, integration of isotropic functions over polygons, and additional figures and tables (DOI: 10.1214/14-AOAS743SUPPB; .pdf).

- Details on likelihood inference for both models.
- Integration of radially symmetric functions over polygonal domains.
- Additional figures and tables of the power-law models for invasive meningococcal disease and influenza.

REFERENCES

ALBERT, R. and BARABÁSI, A.-L. (2002). Statistical mechanics of complex networks. *Rev. Modern Phys.* **74** 47–97. MR1895096

BARTLETT, M. S. (1957). Measles periodicity and community size. *J. Roy. Statist. Soc. Ser. A* **120** 48–70.

BIVAND, R. S., PEBESMA, E. and GÓMEZ-RUBIO, V. (2013). *Applied Spatial Data Analysis with R*, 2nd ed. *Use R!* **10**. Springer, New York. MR3099410

BROCKMANN, D. and HELBING, D. (2013). The hidden geometry of complex, network-driven contagion phenomena. *Science* **342** 1337–1342.

BROCKMANN, D., HUFNAGEL, L. and GEISEL, T. (2006). The scaling laws of human travel. *Nature* **439** 462–465.

BROWN, L. A. and HOLMES, J. (1971). The delimitation of functional regions, nodal regions, and hierarchies by functional distance approaches. *J. Reg. Sci.* **11** 57–72.

BROWN, L. A. and HORTON, F. E. (1970). Functional distance: An operational approach. *Geogr. Anal.* **2** 76–83.

CHILÈS, J.-P. and DELFINER, P. (2012). *Geostatistics*: *Modeling Spatial Uncertainty*, 2nd ed. *Wiley Series in Probability and Statistics* **713**. Wiley, Hoboken, NJ. MR2850475

CHIS STER, I. and FERGUSON, N. M. (2007). Transmission parameters of the 2001 foot and mouth epidemic in Great Britain. *PLoS ONE* **2** e502.

COWLING, B. J., FANG, V. J., RILEY, S., PEIRIS, J. M. S. and LEUNG, G. M. (2009). Estimation of the serial interval of influenza. *Epidemiology* **20** 344–347.

CZADO, C., GNEITING, T. and HELD, L. (2009). Predictive model assessment for count data. *Biometrics* **65** 1254–1261. MR2756513

DAWID, A. P. and SEBASTIANI, P. (1999). Coherent dispersion criteria for optimal experimental design. *Ann. Statist.* **27** 65–81. MR1701101

DEARDON, R., BROOKS, S. P., GRENFELL, B. T., KEELING, M. J., TILDESLEY, M. J., SAVILL, N. J., SHAW, D. J. and WOOLHOUSE, M. E. J. (2010). Inference for individual-level models of infectious diseases in large populations. *Statist. Sinica* **20** 239–261. MR2640693

DIGGLE, P. J. (2006). Spatio-temporal point processes, partial likelihood, foot and mouth disease. *Stat. Methods Med. Res.* **15** 325–336. MR2242245

DIGGLE, P. J. (2007). Spatio-temporal point processes: Methods and applications. In *Statistical Methods for Spatio-Temporal Systems* (B. Finkenstädt, L. Held and V. Isham, eds.) 1–45. Chapman & Hall/CRC, Boca Raton, FL.

DIGGLE, P. J., KAIMI, I. and ABELLANA, R. (2010). Partial-likelihood analysis of spatio-temporal point-process data. *Biometrics* **66** 347–354. MR2758814

ELIAS, J., SCHOULS, L. M., VAN DE POL, I., KEIJZERS, W. C., MARTIN, D. R., GLENNIE, A., OSTER, P., FROSCH, M., VOGEL, U. and VAN DER ENDE, A. (2010). Vaccine preventability of meningococcal clone, Greater Aachen region, Germany. *Emerg. Infect. Dis.* **16** 465–472.

EUBANK, R. L. (2000). Spline regression. In *Smoothing and Regression*: *Approaches*, *Computation*, *and Application* (M. G. Schimek, ed.) 1–18. Wiley, New York.

FAHRMEIR, L. and KNORR-HELD, L. (2000). Dynamic and semiparametric models. In *Smoothing and Regression*: *Approaches*, *Computation*, *and Application* (M. G. Schimek, ed.) 513–544. Wiley, New York.

FANSHAWE, T. R., DIGGLE, P. J., RUSHTON, S., SANDERSON, R., LURZ, P. W. W., GLINIANAIA, S. V., PEARCE, M. S., PARKER, L., CHARLTON, M. and PLESS-MULLOLI, T. (2008). Modelling spatio-temporal variation in exposure to particulate matter: A two-stage approach. *Environmetrics* **19** 549–566. MR2528540

FARRINGTON, C. P., ANDREWS, N. J., BEALE, A. D. and CATCHPOLE, M. A. (1996). A statistical algorithm for the early detection of outbreaks of infectious disease. *J. Roy. Statist. Soc. Ser. A* **159** 547–563. MR1413665

FUHRMANN, C. (2010). The effects of weather and climate on the seasonality of influenza: What we know and what we need to know. *Geography Compass* **4** 718–730.

GEILHUFE, M., HELD, L., SKRØVSETH, S. O., SIMONSEN, G. S. and GODTLIEBSEN, F. (2014). Power law approximations of movement network data for modeling infectious disease spread. *Biom. J.* **56** 363–382.

GIBSON, G. J. (1997). Markov chain Monte Carlo methods for fitting spatiotemporal stochastic models in plant epidemiology. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **46** 215–233.

GNEITING, T., BALABDAOUI, F. and RAFTERY, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **69** 243–268. MR2325275

GNEITING, T. and RAFTERY, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.* **102** 359–378. MR2345548

GNEITING, T. and SCHLATHER, M. (2004). Stochastic models that separate fractal dimension and the Hurst effect. *SIAM Rev.* **46** 269–282 (electronic). MR2114455

GUTENBERG, B. and RICHTER, C. F. (1944). Frequency of earthquakes in California. *Bull. Seismol. Soc. Amer.* **34** 185–188.

HAWKES, A. G. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika* **58** 83–90. MR0278410

HELD, L., HÖHLE, M. and HOFMANN, M. (2005). A statistical framework for the analysis of multivariate infectious disease surveillance counts. *Stat. Model.* **5** 187–199. MR2210732

HELD, L. and PAUL, M. (2012). Modeling seasonality in space–time infectious disease surveillance data. *Biom. J.* **54** 824–843. MR2993630

HELD, L., HOFMANN, M., HÖHLE, M. and SCHMID, V. (2006). A two-component model for counts of infectious diseases. *Biostatistics* **7** 422–437.

HERZOG, S. A., PAUL, M. and HELD, L. (2011). Heterogeneity in vaccination coverage explains the size and occurrence of measles epidemics in German surveillance data. *Epidemiol. Infect.* **139** 505–515.

HÖHLE, M. (2009). Additive-multiplicative regression models for spatio-temporal epidemics. *Biom. J.* **51** 961–978. MR2744450

HÖHLE, M., MEYER, S. and PAUL, M. (2014). *surveillance*: Temporal and spatio-temporal modeling and monitoring of epidemic phenomena. R package version 1.8-0.

HÖHLE, M. and PAUL, M. (2008). Count data regression charts for the monitoring of surveillance time series. *Comput. Statist. Data Anal.* **52** 4357–4368. MR2432467

HÖHLE, M., PAUL, M. and HELD, L. (2009). Statistical approaches to the monitoring and surveillance of infectious diseases for veterinary public health. *Prev. Vet. Med.* **91** 2–10.

HULTH, A., RYDEVIK, G. and LINDE, A. (2009). Web queries as a source for syndromic surveillance. *PLoS ONE* **4** e4378.

JOSSERAN, L., NICOLAU, J., CAILLÈRE, N., ASTAGNEAU, P. and BRÜCKER, G. (2006). Syndromic surveillance based on emergency department activity and crude mortality: Two examples. *Eurosurveillance* **11** 225–229.

KEELING, M. J. and ROHANI, P. (2002). Estimating spatial coupling in epidemiological systems: A mechanistic approach. *Ecol. Lett.* **5** 20–29.

KEELING, M. J. and ROHANI, P. (2008). *Modeling Infectious Diseases in Humans and Animals*. Princeton Univ. Press, Princeton, NJ. MR2354763

KEELING, M. J., WOOLHOUSE, M. E. J., SHAW, D. J., MATTHEWS, L., CHASE-TOPPING, M., HAYDON, D. T., CORNELL, S. J., KAPPEY, J., WILESMITH, J. and GRENFELL, B. T. (2001). Dynamics of the 2001 UK foot and mouth epidemic: Stochastic dispersal in a heterogeneous landscape. *Science* **294** 813–817.

LAWSON, A. B. and LEIMICH, P. (2000). Approaches to the space–time modelling of infectious disease behaviour. *IMA J. Math. Appl. Med. Biol.* **17** 1–13.

LAZER, D., PENTLAND, A., ADAMIC, L., ARAL, S., BARABASI, A.-L., BREWER, D., CHRISTAKIS, N., CONTRACTOR, N., FOWLER, J., GUTMANN, M., JEBARA, T., KING, G., MACY, M., ROY, D. and ALSTYNE, M. V. (2009). Social science. Computational social science. *Science* **323** 721–723.

LE COMBER, S., ROSSMO, D. K., HASSAN, A., FULLER, D. and BEIER, J. (2011). Geographic profiling as a novel spatial tool for targeting infectious disease control. *Int. J. Health Geogr.* **10** 35.

LILJEROS, F., EDLING, C. R., AMARAL, L. A. N., STANLEY, H. E. and ABERG, Y. (2001). The web of human sexual contacts. *Nature* **411** 907–908.

LOMAX, K. S. (1954). Business failures: Another example of the analysis of failure data. *J. Amer. Statist. Assoc.* **49** 847–852.

MEYER, S. (2014). *polyCub*: Cubature over polygonal domains. R package version 0.4-3.

MEYER, S., ELIAS, J. and HÖHLE, M. (2012). A space–time conditional intensity model for invasive meningococcal disease occurrence. *Biometrics* **68** 607–616. MR2959628

MEYER, S. and HELD, L. (2014). Supplement to "Power-law models for infectious disease spread." DOI:10.1214/14-AOAS743SUPPB.

MEYER, S., HELD, L. and HÖHLE, M. (2014). Spatio-temporal analysis of epidemic phenomena using the R package *surveillance*. *J. Stat. Softw.* To appear.

MOSSONG, J., HENS, N., JIT, M., BEUTELS, P., AURANEN, K., MIKOLAJCZYK, R., MASSARI, M., SALMASO, S., TOMBA, G. S., WALLINGA, J., HEIJNE, J., SADKOWSKA-TODYS, M., ROSINSKA, M. and EDMUNDS, W. J. (2008). Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Med.* **5** e74.

NEWMAN, M. E. J. (2005). Power laws, Pareto distributions and Zipf's law. *Contemp. Phys.* **46** 323–351.

NOUFAILY, A., ENKI, D. G., FARRINGTON, P., GARTHWAITE, P., ANDREWS, N. and CHARLETT, A. (2013). An improved algorithm for outbreak detection in multiple surveillance systems. *Stat. Med.* **32** 1206–1222. MR3045892

OGATA, Y. (1998). Space–time point-process models for earthquake occurrences. *Ann. Inst. Statist. Math.* **50** 379–402.

PARETO, V. (1896). *Cours D'Économie Politique* **1**. F. Rouge, Lausanne.

PAUL, M. and HELD, L. (2011). Predictive assessment of a non-linear random effects model for multivariate time series of infectious disease counts. *Stat. Med.* **30** 1118–1136. MR2767846

PAUL, M., HELD, L. and TOSCHKE, A. M. (2008). Multivariate modelling of infectious disease surveillance data. *Stat. Med.* **27** 6250–6267. MR2522320

PINTO, C. M. A., MENDES LOPES, A. and MACHADO, J. A. T. (2012). A review of power laws in real life phenomena. *Commun. Nonlinear Sci. Numer. Simul.* **17** 3558–3578. MR2913993

R CORE TEAM (2013). R: *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

RATTI, C., SOBOLEVSKY, S., CALABRESE, F., ANDRIS, C., READES, J., MARTINO, M., CLAXTON, R. and STROGATZ, S. H. (2010). Redrawing the map of Great Britain from a network of human interactions. *PLoS ONE* **5** e14248.

ROBERT KOCH INSTITUTE (2009). SurvStat@RKI. Available at http://www3.rki.de/SurvStat. Queried on March 6, 2009.

ROSSMO, D. K. (2000). *Geographic Profiling*. CRC Press, Boca Raton.

SCHRÖDLE, B., HELD, L. and RUE, H. (2012). Assessing the impact of a movement network on the spatiotemporal spread of infectious diseases. *Biometrics* **68** 736–744. MR3055178

SOMMARIVA, A. and VIANELLO, M. (2007). Product Gauss cubature over polygons based on Green's integration formula. *BIT* **47** 441–453. MR2334049

SOUBEYRAND, S., HELD, L., HÖHLE, M. and SACHE, I. (2008). Modelling the spread in space and time of an airborne plant disease. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **57** 253–272. MR2440008

TOBLER, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Econ. Geogr.* **46** 234–240.

TRUSCOTT, J., FRASER, C., CAUCHEMEZ, S., MEEYAI, A., HINSLEY, W., DONNELLY, C. A., GHANI, A. and FERGUSON, N. (2012). Essential epidemiological mechanisms underpinning the transmission dynamics of seasonal influenza. *J. R. Soc. Interface* **9** 304–312.

VIBOUD, C., BJØRNSTAD, O. N., SMITH, D. L., SIMONSEN, L., MILLER, M. A. and GRENFELL, B. T. (2006). Synchrony, waves, and spatial hierarchies in the spread of influenza. *Science* **312** 447–451.

WEI, W. and HELD, L. (2014). Calibration tests for count data. *TEST*. DOI:10.1007/s11749-014-0380-8.

WILLEM, L., KERCKHOVE, K. V., CHAO, D. L., HENS, N. and BEUTELS, P. (2012). A nice day for an infection? Weather conditions and social contact patterns relevant to influenza transmission. *PLoS ONE* **7** e48695.

XIE, Y. (2013). *animation*: An R package for creating animations and demonstrating statistical methods. *J. Stat. Softw.* **53** 1–27.

ZIPF, G. K. (1949). *Human Behavior and the Principle of Least Effort*: *An Introduction to Human Ecology*. Addison-Wesley Press, Cambridge, MA.

EPIDEMIOLOGY, BIOSTATISTICS AND PREVENTION INSTITUTE
DEPARTMENT OF BIOSTATISTICS
UNIVERSITY OF ZURICH
HIRSCHENGRABEN 84
CH-8001 ZÜRICH
SWITZERLAND
E-MAIL: Sebastian.Meyer@uzh.ch
Leonhard.Held@uzh.ch
URL: www.biostat.uzh.ch

SUPPLEMENT B OF "POWER-LAW MODELS FOR INFECTIOUS DISEASE SPREAD"*

By Sebastian Meyer and Leonhard Held

University of Zurich

We provide details on likelihood inference in both modelling frameworks, and elaborate on the integration of radially symmetric functions over polygons.

This supplement also contains additional figures and tables of the power-law models for invasive meningococcal disease (IMD) and influenza. Space-time animations are available as a separate supplement (http://www.biostat.uzh.ch/static/powerlaw/).

Note that we refer to equations of the main paper via double parentheses, e.g., ((1)), and to local equations using single parentheses as usual.

CONTENTS

*Funded by the Swiss National Science Foundation (project #137919).

1. Inference details.

1.1. *Individual-level model.* For a point pattern $\{(t_i, \boldsymbol{s}_i) : i = 1, \ldots, n\}$, observed during the period $(0; T]$ in region $\boldsymbol{W}$, the log-likelihood of a point process model, characterized through its conditional intensity function $\lambda(t, \boldsymbol{s})$ with parameter vector $\boldsymbol{\theta}$, is given by

$$l(\boldsymbol{\theta}) = \left[\sum_{i=1}^{n} \log \lambda(t_i, \boldsymbol{s}_i)\right] - \int_0^T \int_{\boldsymbol{W}} \lambda(t, \boldsymbol{s}) \, \mathrm{d}t \, \mathrm{d}\boldsymbol{s} \tag{1}$$

(Daley and Vere-Jones, 2003). See Section 2 for how we solve integration over $\boldsymbol{W}$. Maximization of the log-likelihood is performed numerically by the quasi-Newton algorithm available through the R function `nlminb()`, which is based on FORTRAN 77 routines by Gay (1981).

To find the maximum most efficiently, we make use of the analytical score function and an approximation of the expected Fisher information worked out by Meyer, Elias and Höhle (2012, Web Appendices A and B). Both require the first partial derivatives of the spatial interaction function $f(x)$ with respect to its parameters, which for the power law ((3)) with $\tilde{\sigma} = \log(\sigma)$ and $\tilde{d} = \log(d)$ optimized on the log-scale are as follows:

$$\frac{\partial f(x)}{\partial \tilde{\sigma}} = -d\sigma(x+\sigma)^{-d-1} \qquad \text{and} \qquad \frac{\partial f(x)}{\partial \tilde{d}} = -\frac{\log\left((x+\sigma)^d\right)}{(x+\sigma)^d} \, . \tag{2}$$

For the lagged power law ((4)), we obtain:

$$\frac{\partial f(x)}{\partial \tilde{\sigma}} = \begin{cases} 0, & \text{for } x < \sigma, \\ d\,(\frac{\sigma}{x})^d, & \text{otherwise,} \end{cases}$$

$$\frac{\partial f(x)}{\partial \tilde{d}} = \begin{cases} 0, & \text{for } x < \sigma, \\ -\frac{\log\left((x/\sigma)^d\right)}{(x/\sigma)^d}, & \text{otherwise.} \end{cases}$$

For the Student kernel ((5)), we obtain:

$$\frac{\partial f(x)}{\partial \tilde{\sigma}} = \frac{-2d\sigma^2}{x^2+\sigma^2} \cdot f(x) \qquad \frac{\partial f(x)}{\partial \tilde{d}} = f(x) \cdot \log(f(x))$$

For the step function ((6)) with $\tilde{\alpha}_k = \log(\alpha_k)$, we have

$$\frac{\partial f(x)}{\partial \tilde{\alpha}_k} = \alpha_k \mathbb{1}(x \in I_k) \qquad \text{for } k \in \{1, \ldots, K\} \, .$$

1.2. *Count data model.* Penalized likelihood inference for the count data model ((7)) proceeds by repeating the following two steps until convergence (Paul and Held, 2011):

1. Given $\boldsymbol{\Sigma}$, maximize the penalized log-likelihood $l_{\text{pen}}(\boldsymbol{\theta}, \psi, \boldsymbol{b}; \boldsymbol{\Sigma})$ of regression parameters, where $\boldsymbol{\theta} = (\alpha^{(\lambda)}, \boldsymbol{\beta}^{(\lambda)}, \alpha^{(\phi)}, \boldsymbol{\beta}^{(\phi)}, d, \alpha^{(\nu)}, \boldsymbol{\beta}^{(\nu)})^\top$ covers all fixed effects.
2. Maximize the (approximate) marginal log-likelihood $l_{mar}(\boldsymbol{\Sigma})$ of the covariance matrix $\boldsymbol{\Sigma}$ of $\boldsymbol{b}_i$.

Supplied with the penalized score function and Fisher information matrix, we use the quasi-Newton algorithm available through the R function `nlminb()` to maximize the high-dimensional penalized log-likelihood. However, for the marginal log-likelihood of the variance parameters, we use the Nelder and Mead (1965) algorithm as implemented in `optim()` (based on Pascal code by Nash, 1990), which does not require gradient calculations. It is generally more robust and twice as fast as the previously used quasi-Newton algorithm, since there are only six parameters in $\boldsymbol{\Sigma}$ and evaluation of the marginal score function and Fisher information is complex.

The decay parameter d of the weight function ((9)) or ((10)), respectively, enters the penalized log-likelihood

$$l_{\text{pen}}(\boldsymbol{\theta}, \psi, \boldsymbol{b}; \boldsymbol{\Sigma}) = \sum_{i=1}^{I} \left[\sum_{t=2}^{T} \log f\big(y_{it}; \mu_{it}(\boldsymbol{\theta}, \boldsymbol{b}, \boldsymbol{y}_{t-1}), \psi\big) \right] + \log p(\boldsymbol{b}_i; \boldsymbol{\Sigma})$$

via the mean μ_{it} of f, the probability mass function of the negative binomial distribution. The score function component of d is

$$s(d) = \sum_{i=1}^{I} \sum_{t=2}^{T} \left(\frac{y_{it}}{\mu_{it}} - \frac{\psi^{-1} + y_{it}}{\psi^{-1} + \mu_{it}} \right) \frac{\partial \mu_{it}}{\partial d} \,,$$

where

$$\frac{\partial \mu_{it}}{\partial d} = \phi_{it} \sum_{j \neq i} \frac{\partial w_{ji}}{\partial d} y_{j,t-1}$$

only involves the spatio-temporal component of the model. In addition to the first derivative $\partial w_{ji}/\partial d$ of the weight function, we also need its second derivative to compute the corresponding entry in the Fisher information matrix.

For the raw power-law weights ((9)), first and second derivatives are given by

$$\frac{\partial w_{ji}}{\partial d} = -o_{ji}^{-d} \log o_{ji} \qquad \text{and} \qquad \frac{\partial^2 w_{ji}}{\partial d^2} = o_{ji}^{-d} (\log o_{ji})^2 \,.$$

For normalized power-law weights ((10)) we obtain

$$\frac{\partial w_{ji}}{\partial d} = w_{ji}(d) \cdot \left[-\log o_{ji} - Q_j^{(1)}(d) \right] \quad \text{and}$$
$$\frac{\partial^2 w_{ji}}{\partial d^2} = w_{ji}(d) \cdot \left[\left(-\log o_{ji} - Q_j^{(1)}(d) \right)^2 - Q_j^{(2)}(d) + \left(Q_j^{(1)}(d) \right)^2 \right] ,$$

where

$$Q_j^{(k)}(d) = \frac{\sum_{i=1}^I o_{ji}^{-d} (-\log o_{ji})^k}{\sum_{i=1}^I o_{ji}^{-d}}$$

for $k \in \{1, 2\}$. Estimation of d on the log-scale ($\tilde{d} = \log d$), is performed using

$$\frac{\partial w_{ji}}{\partial \tilde{d}} = \frac{\partial w_{ji}}{\partial d} \cdot d \qquad \text{and} \qquad \frac{\partial^2 w_{ji}}{\partial \tilde{d}^2} = \frac{\partial^2 w_{ji}}{\partial d^2} \cdot d^2 + \frac{\partial w_{ji}}{\partial d} \cdot d \,,$$

which is true for both raw and normalized formulations.

The unconstrained neighbourhood weights ω_o of orders $o \in \{2, \ldots, M\}$ are also estimated on the log-scale: $\tilde{\omega}_o = \log(\omega_o)$. Let $O_{ji}^o := \mathbb{1}(o_{ji} = o)$ for $o \in \{2, \ldots, M-1\}$ and $O_{ji}^M := \mathbb{1}(o_{ji} \geq M)$. For the raw formulation ((11)), first and second partial derivatives are given by

$$\frac{\partial w_{ji}}{\partial \tilde{\omega}_o} = \omega_o \cdot O_{ji}^o \,, \qquad \frac{\partial^2 w_{ji}}{\partial \tilde{\omega}_o^2} = \frac{\partial w_{ji}}{\partial \tilde{\omega}_o} \qquad \text{and} \qquad \frac{\partial^2 w_{ji}}{\partial \tilde{\omega}_o \partial \tilde{\omega}_{o'}} = 0 \text{ for } o \neq o'.$$

For the normalized version $w_{ji}^* := w_{ji} / \sum_{k=1}^I w_{jk}$, we obtain

$$\frac{\partial w_{ji}^*}{\partial \tilde{\omega}_o} = \omega_o \cdot \frac{\Omega_{ji}^o}{\sum_{k=1}^I w_{jk}} \,,$$
$$\frac{\partial^2 w_{ji}}{\partial \tilde{\omega}_o \partial \tilde{\omega}_{o'}} = \frac{\omega_o}{\left(\sum_{k=1}^I w_{jk} \right)^2} \cdot \begin{cases} \Omega_{ji}^o \cdot \left(\sum_{k=1}^I w_{jk} - 2\omega_o S_j^o \right) , & \text{for } o = o', \\ -\omega_{o'} \cdot \left(\Omega_{ji}^{o'} S_j^o + \Omega_{ji}^o S_j^{o'} \right) , & \text{for } o \neq o', \end{cases}$$

where $\Omega_{ji}^o := O_{ji}^o - w_{ji}^* S_j^o$ and $S_j^o := \sum_{k=1}^I O_{jk}^o$.

2. Integration of isotropic functions over polygons. In the spatio-temporal point process model ((1)), evaluation of the log-likelihood (1) and score function involves the spatial integrals

$$\int_{\boldsymbol{W}_i} f_2(\boldsymbol{s})\,\mathrm{d}\boldsymbol{s} = \int_{\boldsymbol{W}_i} f(\|\boldsymbol{s}\|)\,\mathrm{d}\boldsymbol{s}, \qquad i = 1, \ldots, n, \tag{3}$$

of the interaction kernel over shifted versions $\boldsymbol{W}_i := \boldsymbol{W} - \boldsymbol{s}_i$ of the polygonal observation region $\boldsymbol{W} \subset \mathbb{R}^2$ such that the event's location $\boldsymbol{s}_i$ becomes the origin. The score function further requires similar integration of the partial derivatives of f with respect to the kernel parameters, e.g., $\partial f(\|\boldsymbol{s}\|)/\partial\tilde{\sigma}$ and $\partial f(\|\boldsymbol{s}\|)/\partial\tilde{d}$ as listed in Section 1.1 above.

2.1. *Piecewise constant functions, memoization.* Integration is trivial for homogeneous spatial interaction $f(x) \equiv 1$, which assumes mass action kinetics as in simple epidemic models (Keeling and Rohani, 2008). Also for a parametric step kernel ((6)), the above integrals can be solved analytically given a method to calculate areas of intersections of the polygonal integration domain with discs, and thereby calculate the areas of the "rings" of the radially symmetric step function. Note that the intersections are only approximate in that the discs are actually represented by polygons with a high number of vertices. We used 256 vertices and the R package **polyclip** (Johnson, 2013) via **spatstat** (Baddeley and Turner, 2005) to perform polygon intersections.

Additional memoization (using package **memoise** by Wickham, 2010) in computing the "ring areas" for the integration domains $\boldsymbol{W}_i$ proved very useful to reduce computational cost of the step kernel. We generally apply memoization for all spatial interaction functions with respect to calculating the set of integrals (3) for a specific kernel parameter vector (which might be the same in several iterations of log-likelihood maximization). This is especially worthwhile for more general, non-constant spatial interaction functions f, which require cumbersome cubature methods.

2.2. *Gaussian kernel.* For the Gaussian kernel ((2)), the two-dimensional midpoint rule was found to be best suited among several cubature methods if it employs a σ-adaptive bandwidth (Meyer, 2009, Section 3.2). Thereby, an additional trick is applied if a 6σ-circle around $\boldsymbol{s}_i$ – i.e., the numerically "effective range" of interaction – lay completely inside $\boldsymbol{W}$: Then, the spatial integral is approximated by the integral over this circular domain only, which is analytically available for Gaussian f_2 via the χ^2 distribution (Abramowitz and Stegun, 1972, Formula 26.3.24).

However, the simple midpoint rule and the effective range approximation are not appropriate for "spiky" and heavy-tail power-law kernels. Large quantiles for the effective range would rarely fall completely inside $\boldsymbol{W}$ and the midpoint rule would require a very small bandwidth, i.e., a high number of cubature nodes, to accurately catch the peak at $\boldsymbol{s}_i$. Therefore, a more sophisticated cubature method is needed.

2.3. *Product Gauss cubature.* Based on Green's theorem (see Equation (5) in Section 2.4), Sommariva and Vianello (2007a) proposed a cubature method for a continuously differentiable function over a simple closed polygonal domain $\boldsymbol{\Omega} \subset \mathbb{R}^2$ with anticlockwise vertex coordinates $\boldsymbol{v}_0, \boldsymbol{v}_1, \ldots, \boldsymbol{v}_L = \boldsymbol{v}_0$. It is exact for all bivariate polynomials up to a degree $2q-1$, where the number of required cubature nodes depends on the specific shape of the polygon, but is bounded above by $Lq(q+1)$. Their cubature rule incorporates appropriately transformed weights and nodes of one-dimensional Gauss-Legendre quadrature in both dimensions (with q and $q+1$ nodes, respectively), thus the name "product Gauss cubature". It is briefly illustrated in Meyer (2009, Section 3.2.5) and available either in its original MATLAB implementation by Sommariva and Vianello (2007b) or as an R port in package **polyCub** (Meyer, 2014).

However, we can greatly improve the efficiency of numerical integration by taking analytical advantage of the assumed isotropy of spatial interaction. Specifically, we will have to apply numerical integration in only one dimension – a line integral along the polygon boundary – with the integrand being available in closed-form.

2.4. *Efficient integration using Green's theorem.* The path $\gamma\colon [0,1] \to \mathbb{R}^2$ is defined to parametrize a positively oriented, piecewise smooth, simple closed curve describing the boundary of a region $\boldsymbol{D} \subset \mathbb{R}^2$. Further be $f_2\colon \mathbb{R}^2 \to \mathbb{R}$ a radially symmetric, continuously differentiable function with $(x,y) \mapsto f(\|(x,y)\|)$. Then,

$$\iint_{\boldsymbol{D}} f_2(x,y)\,\mathrm{d}x\,\mathrm{d}y = \int_0^1 \frac{\langle \gamma'(t), \hat{\gamma}(t)\rangle}{\|\gamma(t)\|^2} F(\|\gamma(t)\|)\,\mathrm{d}t\,, \tag{4}$$

where

$$\gamma'(t) = \begin{pmatrix} x'(t) \\ y'(t) \end{pmatrix}, \qquad \hat{\gamma}(t) = \begin{pmatrix} -y(t) \\ x(t) \end{pmatrix},$$

and $F\colon \mathbb{R}_0^+ \to \mathbb{R}$ is defined by

$$F(R) = \int_0^R r f(r)\,\mathrm{d}r\,.$$

Equation (4) can be proved using Green's theorem

$$\iint_{\boldsymbol{D}} \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) \mathrm{d}x \, \mathrm{d}y = \oint_{\partial \boldsymbol{D}} P(x,y) \, \mathrm{d}x + Q(x,y) \, \mathrm{d}y \tag{5}$$

with

$$P(x,y) = \frac{-yF(\|(x,y)\|)}{\|(x,y)\|^2} \qquad \text{and} \qquad Q(x,y) = \frac{xF(\|(x,y)\|)}{\|(x,y)\|^2} \,,$$

and we thank Emil Hedevang and Christian Reiher for valuable support in ascertaining this useful integration formula.

To apply Equation (4) for the special case of a polygonal region $\boldsymbol{D} = \boldsymbol{\Omega}$, we conveniently define the boundary curve γ piecewise in that the lth polygon edge $[\boldsymbol{v}_{l-1}, \boldsymbol{v}_l]$, $l \in \{1, \ldots, L\}$, is described by the path

$$\gamma_l(t) := \boldsymbol{v}_{l-1} + t \cdot (\boldsymbol{v}_l - \boldsymbol{v}_{l-1}) \tag{6}$$

for $t \in [0,1]$. Hence, Equation (4) becomes

$$\iint_{\boldsymbol{\Omega}} f_2(x,y) \, \mathrm{d}x \, \mathrm{d}y = \sum_{l=1}^{L} \int_0^1 \frac{\langle \boldsymbol{v}_l, \hat{\boldsymbol{v}}_{l-1} \rangle}{\|\gamma_l(t)\|^2} F(\|\gamma_l(t)\|) \, \mathrm{d}t \,, \tag{7}$$

where the "hat"-notation of a vertex ($\hat{\boldsymbol{v}}$) means coordinate switch from $(x,y)^\top$ to $(-y,x)^\top$ as for $\hat{\gamma}(t)$.

In practice, Equation (7) turns into a cubature rule by using numerical integration along the paths γ_l. We employ the adaptive Gauss-Kronrod quadrature method provided by the R function `integrate()`, which is based on QUADPACK routines by Piessens et al. (1983).

The presented isotropic cubature method is implemented in the R package **polyCub** (Meyer, 2014). It is mainly useful for spatial interaction functions f which allow a closed-form expression of the integral function F. If this is not feasible, F can either be evaluated numerically using `integrate()` as well, or the product Gauss cubature can be applied. For the various power-law kernels, F can be solved analytically which results in tremendous gain of speed and accuracy. For instance, fitting the power-law model for the IMD data using product Gauss cubature with $q = 40$ took 80 minutes, whereas it took only half this time using isotropic cubature – or 11 minutes if additionally parallelized on 4 cores with respect to i in (3).

The analytical solution of $F(R)$ for the power law $f(x) = (x+\sigma)^{-d}$ is

$$\begin{aligned} F(R;\sigma,d) &= \int_0^R x(x+\sigma)^{-d}\,\mathrm{d}x \\ &= \begin{cases} R - \sigma\log(\frac{R}{\sigma}+1), & d=1, \\ \log(\frac{R}{\sigma}+1) - \frac{R}{R+\sigma}, & d=2, \\ \frac{1}{1-d}\left(R(R+\sigma)^{1-d} - \frac{(R+\sigma)^{2-d}-\sigma^{2-d}}{2-d}\right), & \text{otherwise.} \end{cases} \end{aligned}$$

The isotropic cubature method is equally applicable to the partial derivatives of f listed in Section 1.1 above, which need to be integrated as part of the score function over the same polygonal domains $\boldsymbol{W}_i$ as f itself. For instance, for the partial derivatives (2) of the power law, we obtain

$$F_{\tilde{\sigma}}(R;\sigma,d) = \int_0^R x\frac{\partial f(x)}{\partial\tilde{\sigma}}\,\mathrm{d}x = -d\sigma\int_0^R x(x+\sigma)^{-d-1}\,\mathrm{d}x = -d\sigma F(R;\sigma,d+1)$$

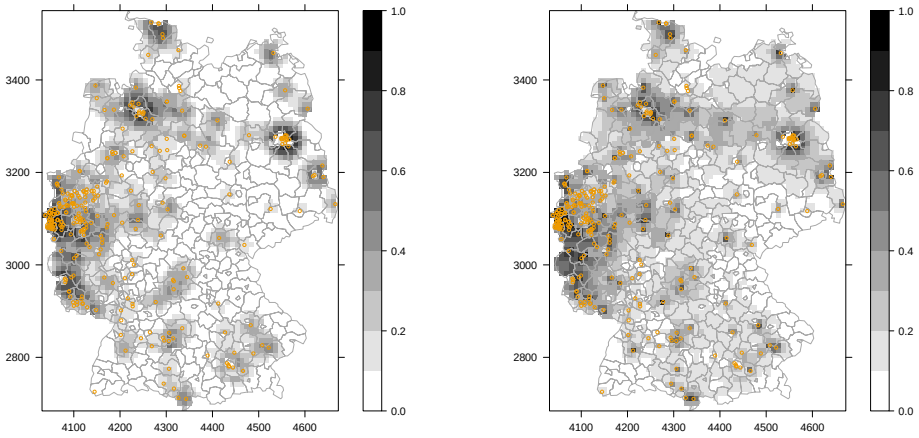
and

$$F_{\tilde{d}}(R;\sigma,d) = \int_0^R x\frac{\partial f(x)}{\partial\tilde{d}}\,\mathrm{d}x = \int_0^R xf(x)\log(f(x))\,\mathrm{d}x$$

$$= \begin{cases} \sigma\log\sigma(1-\frac{\log\sigma}{2}) - (R+\sigma)\log(R+\sigma) + \frac{\sigma}{2}(\log(R+\sigma))^2 + R, & d=1, \\ \frac{1}{R+\sigma}(-\log(R+\sigma)((R+\sigma)\log(R+\sigma)+2\sigma) + (R+\sigma)\log\sigma(\log\sigma+2)+2R), & d=2, \\ d\frac{\sigma^{2-d}\left((3d-d^2-2)\log\sigma-2d+3\right)+(R+\sigma)^{1-d}\left(\log(R+\sigma)(d-1)(d-2)(R(d-1)+\sigma)+R(d^2+1)+2d(\sigma-R)-3\sigma\right)}{(d-1)^2(d-2)^2}, & \text{else.} \end{cases}$$

3. Additional figures and tables.

3.1. *IMD models.*



(a) Gaussian kernel, B-type.

(b) Power-law kernel, B-type.

(c) Gaussian kernel, C-type.

(d) Power-law kernel, C-type.

FIG 1. *Maps showing the epidemic proportion of the accumulated total intensity, formally* $1-\int_0^T \hat{\nu}_{[t][\boldsymbol{s}]}\,\rho_{[t][\boldsymbol{s}]}\,\mathrm{d}t/\int_0^T \hat{\lambda}(t,\boldsymbol{s})\,\mathrm{d}t$*, separately for both finetypes and both spatial interaction functions* ((2)) *and* ((3))*, respectively.*

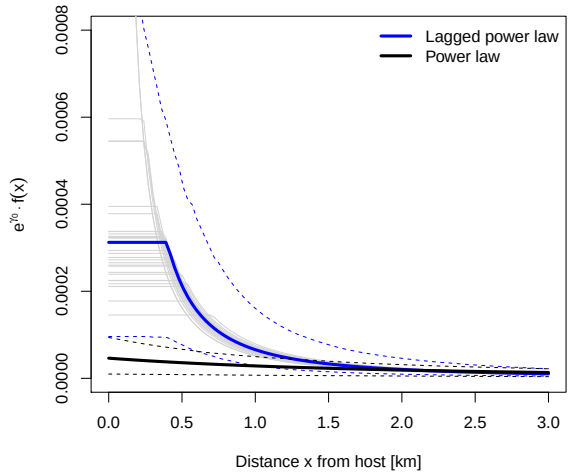


FIG 2. *Estimated lagged power-law kernel* ((4)) *scaled by the epidemic intercept* $\exp(\gamma_0)$. *For comparison, the plot also shows the estimated power law as in Figure 3 of the main paper (but note the very different axis ranges). Similarly, the dashed lines represent 95% bootstrap confidence intervals and the light grey lines are estimates obtained from the sensitivity analysis with repeated random tie-breaking. This shows that the uniform short-range dispersal radius* σ *of the lagged power law is not well identified.*

TABLE 1

Parameter estimates and 95% Wald confidence intervals for the models with Student, lagged power-law, and step function kernel, respectively. Estimates for the Gaussian and the power-law model are shown in Table 1 of the main paper. Note that σ, d *and* γ_0 *are not directly comparable; instead see Figure 2 above for the lagged power law, and Figure 3 of the main paper.*

	Lagged power law ((4))		Student ((5))		Step function ((6))	
	Estimate	95% CI	Estimate	95% CI	Estimate	95% CI
β_0	-20.58	-20.69 to -20.47	-20.57	-20.68 to -20.47	-20.58	-20.68 to -20.48
β_{trend}	-0.04	-0.09 to 0.00	-0.05	-0.09 to -0.00	-0.05	-0.09 to 0.00
β_{sin}	0.25	0.12 to 0.39	0.26	0.12 to 0.39	0.25	0.12 to 0.39
β_{cos}	0.27	0.13 to 0.40	0.27	0.14 to 0.40	0.27	0.14 to 0.41
γ_0	-8.07	-9.22 to -6.92	-6.61	-9.00 to -4.22	-9.93	-10.72 to -9.13
γ_{C}	-0.72	-1.25 to -0.19	-0.80	-1.30 to -0.29	-0.86	-1.35 to -0.36
$\gamma_{\text{3-18}}$	0.87	0.10 to 1.64	0.75	0.09 to 1.41	0.69	0.06 to 1.32
$\gamma_{\geq 19}$	-0.33	-1.48 to 0.82	-0.17	-1.06 to 0.73	-0.38	-1.32 to 0.56
σ	0.40	0.18 to 0.86	6.70	3.78 to 11.88		
d	1.69	1.52 to 1.88	1.21	0.92 to 1.60		
α_1					0.06	0.01 to 0.56
α_2					0.12	0.05 to 0.29
α_3					0.04	0.02 to 0.10
α_4					0.02	0.01 to 0.05
α_5					0.01	0.00 to 0.01
α_6					0.00	0.00 to 0.00

3.2. *Influenza models.*

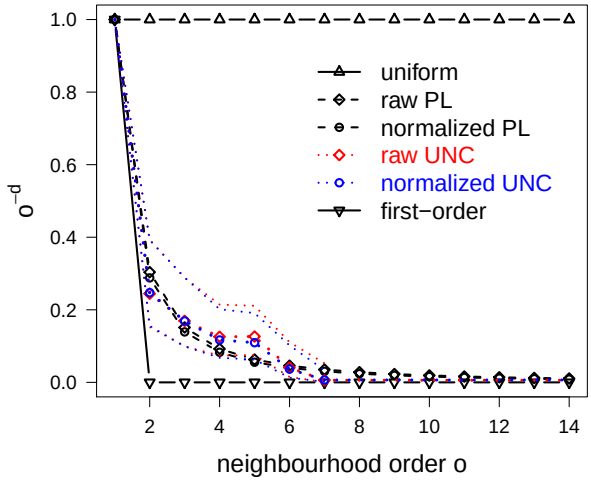


FIG 3. *Estimated power law ((9, PL)) and unconstrained weights ((11, UNC)) with 95% confidence intervals. This plot shows the raw versions for comparison with the normalized ones already included in Figure 5a of the main paper. See Table 2 below for associated parameter estimates.*

TABLE 2

Estimated parameters (with standard errors) excluding intercepts and trend/seasonal coefficients for the models with unconstrained neighbourhood weights. For comparison with the other models, see Table 3 in the main paper.

	raw	normalized
ω_2	0.24 (0.06)	0.25 (0.06)
ω_3	0.17 (0.05)	0.17 (0.05)
ω_4	0.13 (0.03)	0.12 (0.03)
ω_5	0.13 (0.03)	0.11 (0.03)
ω_6	0.04 (0.02)	0.04 (0.02)
ω_7	0.01 (0.01)	0.01 (0.01)
ψ	0.85 (0.03)	0.85 (0.03)
σ^2_λ	0.17	0.16
σ^2_ϕ	0.96	0.92
σ^2_ν	0.67	0.67
$\rho_{\lambda\phi}$	0.2	0.21
$\rho_{\lambda\nu}$	0.32	0.3
$\rho_{\phi\nu}$	0.32	0.32
$l_{\text{pen}}(l_{\text{mar}})$	-18118 (-462)	-18114 (-458)

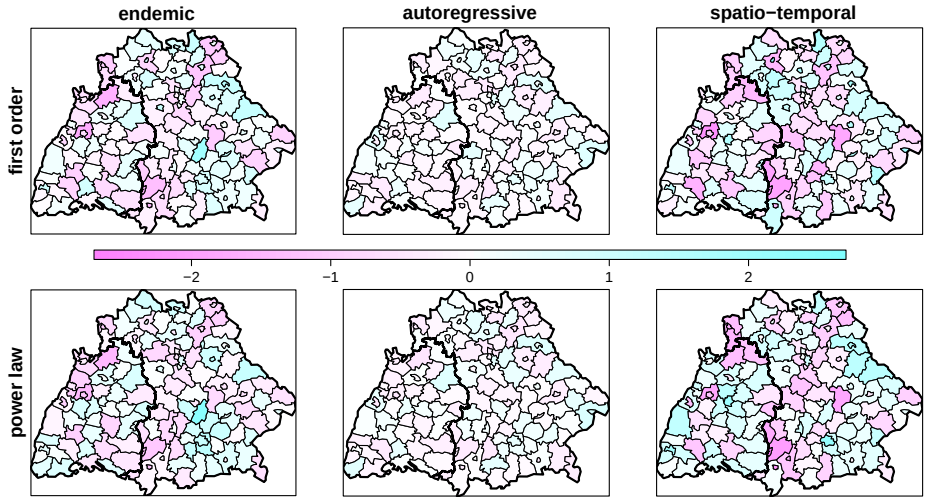


FIG 4. *Maps of the estimated district- and component-specific random effects (in models with normalized weights).*

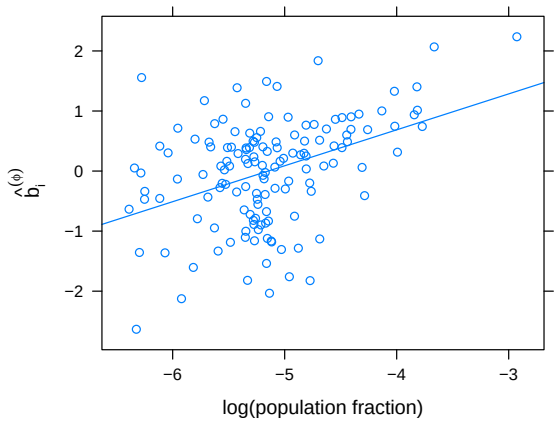


FIG 5. *Scatterplot of $\log(pop_i)$ against the estimated random effects $\hat{b}_i^{(\phi)}$ from the normalized power-law model without explicit account of the district population in ϕ_{it}. The corresponding correlation coefficient is 0.41.*

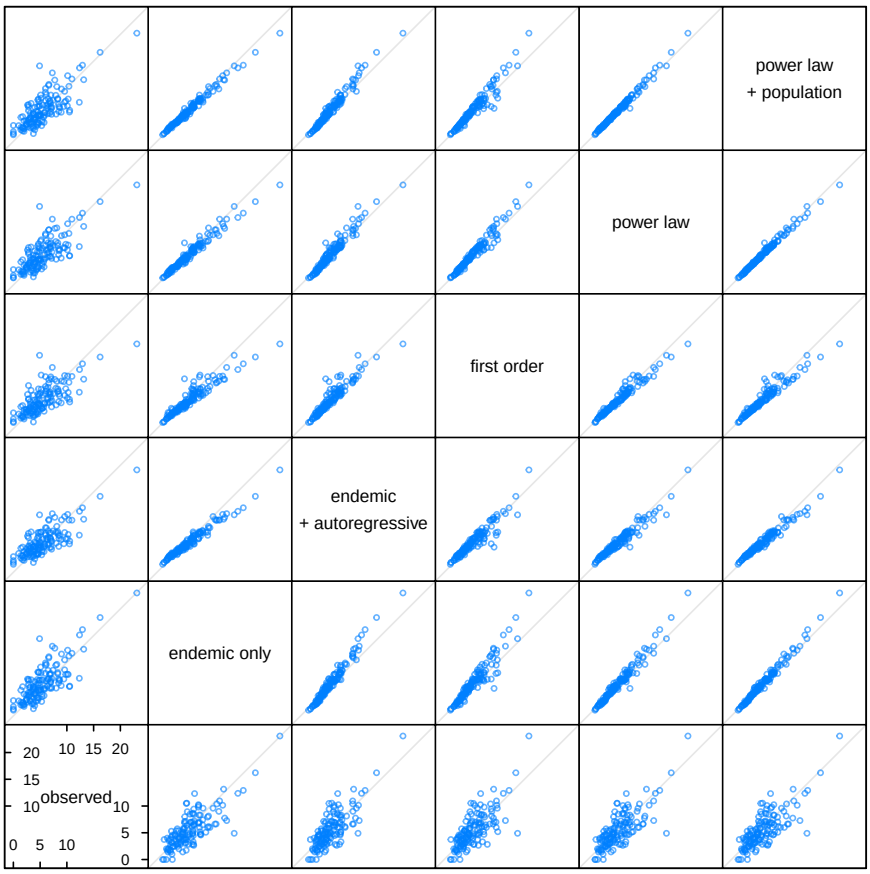


FIG 6. *Observed vs. mean simulated number of cases ($\sqrt{}$-transformed!) during the 2008 wave of influenza in the 140 administrative districts of Bavaria and Baden-Württemberg as predicted by 5 different models.*

References.

Abramowitz, M. and Stegun, I. A., eds. (1972). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. National Bureau of Standards Applied Mathematics Series*. Dover Publications, New York.

Baddeley, A. and Turner, R. (2005). **spatstat**: An R package for analyzing spatial point patterns. *Journal of Statistical Software* **12** 1–42.

Daley, D. J. and Vere-Jones, D. (2003). *An Introduction to the Theory of Point Processes*, 2nd ed. *Probability and its Applications* **I: Elementary Theory and Methods**. Springer-Verlag, New York.

Gay, D. M. (1981). Computing optimal locally constrained steps. *SIAM Journal on Scientific and Statistical Computing* **2** 186–197.

Green, G. (1828). *An Essay on the Application of Mathematical Analysis to the Theories of Electricity and Magnetism*. T. Wheelhouse, Nottingham.

Johnson, A. (2013). **polyclip**: Polygon Clipping. R package version 1.2-0, ported to R by Adrian Baddeley and Brian Ripley.

Keeling, M. J. and Rohani, P. (2008). *Modeling Infectious Diseases in Humans and Animals*. Princeton University Press.

Meyer, S. (2009). Spatio-Temporal Infectious Disease Epidemiology based on Point Processes Master's Thesis, Department of Statistics, Ludwig-Maximilians-Universität, München.

Meyer, S. (2014). **polyCub**: Cubature over Polygonal Domains. R package version 0.4-2.

Meyer, S., Elias, J. and Höhle, M. (2012). A space-time conditional intensity model for invasive meningococcal disease occurrence. *Biometrics* **68** 607–616.

Nash, J. C. (1990). *Compact Numerical Methods for Computers: Linear Algebra and Function Minimisation*, 2nd ed. Adam Hilger.

Nelder, J. A. and Mead, R. (1965). A simplex method for function minimization. *The Computer Journal* **7** 308–313.

Paul, M. and Held, L. (2011). Predictive assessment of a non-linear random effects model for multivariate time series of infectious disease counts. *Statistics in Medicine* **30** 1118–1136.

Piessens, R., de Doncker-Kapenga, E., Überhuber, C. W. and Kahaner, D. K. (1983). *QUADPACK: A Subroutine Package for Automatic Integration. Springer Series in Computational Mathematics* **1**. Springer.

Sommariva, A. and Vianello, M. (2007a). Product Gauss cubature over polygons based on Green's integration formula. *Bit Numerical Mathematics* **47** 441–453.

Sommariva, A. and Vianello, M. (2007b). `polygauss`: MATLAB code for Gauss-like cubature over polygons. `http://www.math.unipd.it/~alvise/software.html`.

R Core Team (2013). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Wickham, H. (2010). **memoise**: Memoise functions. R package version 0.1.

University of Zurich
Institute of Social and Preventive Medicine
Division of Biostatistics
Hirschengraben 84
CH-8001 Zürich
Switzerland
E-mail: Sebastian.Meyer@ifspm.uzh.ch
Leonhard.Held@ifspm.uzh.ch
URL: www.biostat.uzh.ch

PAPER II

Spatio-temporal analysis of epidemic phenomena using the R package `surveillance`

Sebastian Meyer, Leonhard Held, Michael Höhle

In press at the *Journal of Statistical Software*, 2016.

Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. *doi: 10.18637/jss.v000.i00*

Spatio-Temporal Analysis of Epidemic Phenomena Using the R Package surveillance

Sebastian Meyer
University of Zurich

Leonhard Held
University of Zurich

Michael Höhle
Stockholm University

Abstract

The availability of geocoded health data and the inherent temporal structure of communicable diseases have led to an increased interest in statistical models and software for spatio-temporal data with epidemic features. The open source R package **surveillance** can handle various levels of aggregation at which infective events have been recorded: individual-level time-stamped geo-referenced data (case reports) in either continuous space or discrete space, as well as counts aggregated by period and region. For each of these data types, the **surveillance** package implements tools for visualization, likelihoood inference and simulation from recently developed statistical regression frameworks capturing endemic and epidemic dynamics. Altogether, this paper is a guide to the spatio-temporal modeling of epidemic phenomena, exemplified by analyses of public health surveillance data on measles and invasive meningococcal disease.

Keywords: spatio-temporal surveillance data, endemic-epidemic modeling, infectious disease epidemiology, self-exciting point process, multivariate time series of counts, branching process with immigration.

1. Introduction

Epidemic data are realizations of spatio-temporal processes with autoregressive or "self-exciting" behavior. Examples of epidemic phenomena beyond infectious diseases include earth quakes (Ogata 1999), crimes (Johnson 2010; Mohler, Short, Brantingham, Schoenberg, and Tita 2011), invasive species (Balderama, Schoenberg, Murray, and Rundel 2012), and forest fires (Vrbik, Deardon, Feng, Gardner, and Braun 2012). Epidemic data are special with regard to at least three aspects, which hinder the application of classical statistical approaches: the data are rarely a result of planned experiments, the observations (cases, events) are not independent, and often the process is only partially observable.

Since 2005, the open source R (R Core Team 2015) package **surveillance** provides a growing computational framework for methodological developments and practical tools for the *monitoring* and *modeling* of epidemic phenomena – traditionally in the context of infectious diseases. Monitoring is concerned with prospective aberration detection for which several algorithms have been implemented as described by Höhle (2007) and recently updated and reviewed by Salmon, Schumacher, and Höhle (2016). The other major purpose of the **surveillance** package and the focus of this paper is the regression-oriented modeling of spatio-temporal epidemic data. This enables the user to a) assess the role of environmental factors, socio-demographic characteristics, or control measures in shaping endemic and epidemic dynamics, b) analyze the spatio-temporal interaction of events, and c) simulate the epidemic spread from estimated models.

The implemented statistical modeling frameworks have already been successfully applied to a broad range of surveillance data, e.g., human influenza (Paul, Held, and Toschke 2008; Paul and Held 2011; Geilhufe, Held, Skrøvseth, Simonsen, and Godtliebsen 2014), meningococcal disease (Paul *et al.* 2008; Paul and Held 2011; Meyer, Elias, and Höhle 2012), measles (Herzog, Paul, and Held 2011), psychiatric hospital admissions (Meyer, Warnke, Rössler, and Held 2015), rabies in foxes (Höhle, Paul, and Held 2009), coxiellosis in cows (Schrödle, Held, and Rue 2012), and the classical swine fever virus (Höhle 2009). Although these applications all originate from public or animal health surveillance, we stress that our methods also apply to the other epidemic phenomena described above.

To the best of our knowledge, no other software can estimate regression models for spatio-temporal epidemic data. There are, however, some related R packages that we like to mention here, since they also deal with epidemic phenomena. For instance, the *R-epi project*[1] lists the package **EpiEstim** (Cori, Ferguson, Fraser, and Cauchemez 2013), which can estimate the average number of secondary cases caused by an infected individual, the so-called reproduction number, from a time series of disease incidence. Similar functionality is provided by the package **R0** (Obadia, Haneef, and Boelle 2012). Other packages are designed to estimate transmission characteristics from phylogenetic trees (**TreePar**, Stadler and Bonhoeffer 2013), or to reconstruct transmission trees from sequence data (**outbreaker**, Jombart, Cori, Didelot, Cauchemez, Fraser, and Ferguson 2014). The package **amei** (Merl, Johnson, Gramacy, and Mangel 2010) is targeted towards finding optimal intervention strategies, e.g., the proportion of the population to be vaccinated to prevent further disease spread, using purely temporal epidemic models. The recently published package **tscount** (Liboschik, Fokianos, and Fried 2015) is dedicated to the analysis of count time series with serial correlation such as the number of stock market transitions per minute or the weekly number of reported infections of a particular disease. The **tscount** package can fit a univariate version of the areal count time-series model presented in Section 5. For a purely spatial analysis of disease occurrence, see, e.g., the recent paper by Brown (2015) introducing the package **diseasemapping**. One of the few packages fitting spatio-temporal epidemic models is **etasFLP** (Adelfio and Chiodi 2015). The Epidemic-Type Aftershock-Sequences (ETAS) model for earthquakes (Ogata 1999) is closely related to the endemic-epidemic point process model described in Section 3, but incorporates seismological laws rather than covariates. The long-standing package **splancs** (Rowlingson and Diggle 2015) offers diagnostic tools to investigate space-time clustering in a point pattern, i.e., to check if the process at hand shows self-exciting epidemic behavior. Statistical tests for space-time interaction are discussed in Meyer *et al.* (2015), who propose

[1] `https://sites.google.com/site/therepiproject/`

a test based on the regression framework of Section 3. An important recent development for spatio-temporal tasks in R are the basic data classes and utility functions provided by the dedicated package **spacetime** (Pebesma 2012), which builds upon the quasi standards **sp** (Bivand, Pebesma, and Gómez-Rubio 2013) for spatial data and **xts** (Ryan and Ulrich 2014) for time-indexed data, respectively. For a more general overview of R packages for spatio-temporal data, see the CRAN Task View "Handling and Analyzing Spatio-Temporal Data" (Pebesma 2016). A non-R option is the *Spatiotemporal Epidemiological Modeler* (STEM) tool[2]. It has a graphical user interface and can simulate the evolution of disease incidence in a population. The ability to estimate model parameters from surveillance data, however, is limited to simple non-spatial models. WinBUGS has been used for Bayesian inference of specialized spatio-temporal epidemic models (Malesios, Demiris, Kalogeropoulos, and Ntzoufras 2014).

The remainder of this paper is organized as follows: Section 2 gives a brief overview of the three statistical models for spatio-temporal epidemic data implemented in **surveillance**. Each of the subsequent model-specific Sections 3 to 5 first describes the associated methodology and then illustrates the model implementation – including data handling, visualization, inference, and simulation – by applications to infectious disease surveillance data. Section 6 concludes the paper.

2. Spatio-temporal endemic-epidemic modeling

Epidemic models traditionally describe the spread of a communicable disease in a population. Often, a compartmental view of the population is taken, placing individuals into one of the three states (S)usceptible, (I)nfectious, or (R)emoved. Modeling the transitions between these states in a closed population using deterministic differential equations dates back to the work of Kermack and McKendrick (1927). Considering a stochastic version of the simplest homogeneous SIR model in a closed population of size N, the hazard rate for a susceptible individual $i \in S(t)$ to become infectious at time t – the so-called force of infection – is

$$\lambda_i(t) = \sum_{j \in I(t)} \beta \,. \tag{1}$$

Here, $S(t), I(t) \subseteq \{1, \ldots, N\}$ denote the index sets of currently susceptible and infectious individuals, respectively, and the parameter $\beta > 0$ is called the transmission rate. The stochastic SIR model is complemented by a distributional assumption about how long individuals are infective, where typical choices are the exponential or the gamma distribution. The set of recovered individuals at time t is found as $R(t) = \{1, \ldots, N\} \setminus (S(t) \cup I(t))$. The above homogeneous SIR model has since been extended in a multitude of ways, e.g., by additional states (addressing population heterogeneities arising from age groups, spatial location or vaccination) or population demographics. Overviews of SIR modeling approaches can be found in Anderson and May (1991), Daley and Gani (1999), and Keeling and Rohani (2008). The estimation of SIR model parameters from actual observed data is, however, often only treated marginally in such descriptions. In contrast, a number of more statistically flavored epidemic models have emerged recently. This includes, e.g., the TSIR model (Finkenstädt and Grenfell 2000), two-component time-series models (Held, Höhle, and Hofmann 2005; Held, Hofmann,

[2] `http://www.eclipse.org/stem/`

Höhle, and Schmid 2006), and point process models (Lawson and Leimich 2000; Diggle 2006). An overview of temporal and spatio-temporal epidemic models and their relation to the underlying metapopulation SIR models can be found in Höhle (2016).

At the heart of any statistical analysis is the subject-matter scientific problem, which a data-driven analysis seeks to address. Due to the generality and complexity of such problems we adopt here a technocratic view and let the available data guide what a "useful" epidemic model is. The **surveillance** package offers regression-oriented modeling frameworks for three different types of spatio-temporal data distinguished by the spatial and temporal resolution (Table 1). First, if an entire region is continuously monitored for infective events, which are time-stamped, geo-referenced, and potentially enriched with further event-specific data, then a (marked) spatio-temporal point pattern arises. Such continuous space-time epidemic data can be viewed as a realization of a self-exciting spatio-temporal point process (Section 3). The second data type we consider comprises the event history of a discrete set of units followed over time – e.g., farms during livestock epidemics – while registering when they become susceptible, infected, and potentially removed (neither at risk nor infectious). These data fit into the framework of a spatial SIR model represented as a multivariate temporal point process (Section 4). Our third data type is often encountered as a result of privacy protection or reporting regimes, and is an aggregated version of the individual event data mentioned first: event counts by region and period. Such areal count time series can be fitted with the multivariate negative binomial time-series model presented in Section 5.

The three aforementioned model classes are all inspired by the Poisson branching process with immigration approach proposed by Held *et al.* (2005). Its main characteristic is the additive decomposition of disease risk into *endemic* and *epidemic* features, similar to the *background* and *triggered* components in the ETAS model for earthquake occurrence. The endemic component describes the risk of new events by external factors independent of the history of the epidemic process. In the context of infectious diseases, such factors may include seasonality, population density, socio-demographic variables, and vaccination coverage – all potentially varying in time and/or space. Explicit dependence between events is then introduced through an epidemic component driven by the observed past.

Each of the following three model-specific sections starts with a brief theoretical introduction to the respective spatio-temporal endemic-epidemic model, before we describe the implementation using the example data mentioned in Table 1.

	`twinstim` **(Section 3)**	`twinSIR` **(Section 4)**	`hhh4` **(Section 5)**
Data class	`epidataCS`	`epidata`	`sts`
Resolution	individual events in continuous space-time	individual SI[R][S] event history of a fixed population	event counts aggregated by region and time period
Example	cases of meningococcal disease, Germany, 2002–8	measles outbreak among children in Hagelloch, 1861	weekly counts of measles by district, Weser-Ems, 2001–2
Model	(marked) spatio-temporal point process	multivariate temporal point process	multivariate time series (Poisson or NegBin)
Reference	Meyer *et al.* (2012)	Höhle (2009)	Held and Paul (2012)

Table 1: Spatio-temporal endemic-epidemic models and corresponding data classes implemented in the R package **surveillance**.

3. Spatio-temporal point pattern of infective events

The endemic-epidemic spatio-temporal point process model "`twinstim`" is designed for point-referenced, individual-level surveillance data. As an illustrative example, we use case reports of invasive meningococcal disease (IMD) caused by the two most common bacterial finetypes of meningococci in Germany, 2002–2008, as previously analyzed by Meyer *et al.* (2012) and Meyer and Held (2014a). We start by describing the general model class in Section 3.1. Section 3.2 introduces the example data and the associated class `epidataCS`, Section 3.3 presents the core functionality of fitting and analyzing such data using `twinstim`, and Section 3.4 shows how to simulate realizations from a fitted model.

3.1. Model class: `twinstim`

Infective events occur at specific points in continuous space and time, which gives rise to a spatio-temporal point pattern $\{(\boldsymbol{s}_i, t_i) : i = 1, \ldots, n\}$ from a region $\boldsymbol{W}$ observed during a period $(0, T]$. The locations $\boldsymbol{s}_i$ and time points t_i of the n events can be regarded as a realization of a self-exciting spatio-temporal point process, which can be characterized by its conditional intensity function (CIF, also termed intensity process) $\lambda(\boldsymbol{s}, t)$. It represents the instantaneous event rate at location $\boldsymbol{s}$ at time point t given all past events, and is often more verbosely denoted by λ^* or by explicit conditioning on the "history" $\mathcal{H}_t$ of the process. Daley and Vere-Jones (2003, Chapter 7) provide a rigorous mathematical definition of this concept, which is key to likelihood analysis and simulation of "evolutionary" point processes.

Meyer *et al.* (2012) formulated the model class "`twinstim`" – a *two*-component *s*patio-*t*emporal *i*ntensity *m*odel – by a superposition of an endemic and an epidemic component:

$$\lambda(\boldsymbol{s}, t) = \nu_{[\boldsymbol{s}][t]} + \sum_{j \in I(\boldsymbol{s},t)} \eta_j \, f(\|\boldsymbol{s} - \boldsymbol{s}_j\|) \, g(t - t_j) \, . \tag{2}$$

This model constitutes a branching process with immigration. Part of the event rate is due to the first, endemic component, which reflects sporadic events caused by unobserved sources of infection. This background rate of new events is modelled by a log-linear predictor $\nu_{[\boldsymbol{s}][t]}$ incorporating regional and/or time-varying characteristics. Here, the space-time index $[\boldsymbol{s}][t]$ refers to the region covering $\boldsymbol{s}$ during the period containing t and thus spans a whole spatio-temporal grid on which the involved covariates are measured, e.g., district $\times$ month. We will later see that the endemic component therefore simply equals an inhomogeneous Poisson process for the event counts by cell of that grid.

The second, observation-driven epidemic component adds "infection pressure" from the set

$$I(\boldsymbol{s}, t) = \{j : t_j < t \ \wedge \ t - t_j \leq \tau_j \ \wedge \ \|\boldsymbol{s} - \boldsymbol{s}_j\| \leq \delta_j\}$$

of past events and hence makes the process "self-exciting". During its infectious period of length τ_j and within its spatial interaction radius δ_j, the model assumes each event j to trigger further events, which are called offspring, secondary cases, or aftershocks, depending on the application. The triggering rate (or force of infection) is proportional to a log-linear predictor η_j associated with event-specific characteristics ("marks") $\boldsymbol{m}_j$, which are usually attached to the point pattern of events. The decay of infection pressure with increasing spatial and temporal distance from the infective event is modelled by parametric interaction functions f and g, respectively (Lawson and Leimich 2000, Section 4). A simple assumption

for the time course of infectivity is $g(t) = 1$. Alternatives include exponential decay, a step function, or empirically derived functions such as Omori's law for aftershock intervals (Utsu, Ogata, and Matsu'ura 1995). With regard to spatial interaction, a Gaussian kernel $f(x) = \exp\{-x^2/(2\sigma^2)\}$ could be chosen. However, in modeling the spread of human infectious diseases on larger scales, a heavy-tailed power-law kernel $f(x) = (x + \sigma)^{-d}$ was found to perform better (Meyer and Held 2014a). The (possibly infinite) upper bounds τ_j and δ_j provide a way of modeling event-specific interaction ranges. However, since these need to be pre-specified, a common assumption is $\tau_j \equiv \tau$ and $\delta_j \equiv \delta$, where the infectious period τ and the spatial interaction radius δ are determined by subject-matter considerations.

Model-based effective reproduction numbers

Similar to the simple SIR model (see, e.g., Keeling and Rohani 2008, Section 2.1), the above point process model (2) features a reproduction number derived from its branching process interpretation. As soon as an event occurs (individual becomes infected), it triggers offspring (secondary cases) around its origin $(\boldsymbol{s_j}, t_j)$ according to an inhomogeneous Poisson process with rate $\eta_j\, f(\|\boldsymbol{s} - \boldsymbol{s_j}\|)\, g(t - t_j)$. Since this triggering process is independent of the event's parentage and of other events, the expected number μ_j of events triggered by event j can be obtained by integrating the triggering rate over the observed interaction domain:

$$\mu_j = \eta_j \cdot \left[\int_0^{\min(T-t_j,\tau_j)} g(t)\, dt\right] \cdot \left[\int_{\boldsymbol{R}_j} f(\|\boldsymbol{s}\|)\, d\boldsymbol{s}\right] , \tag{3}$$

where

$$\boldsymbol{R}_j = (b(\boldsymbol{s}_j, \delta_j) \cap \boldsymbol{W}) - \boldsymbol{s}_j \tag{4}$$

is event j's influence region centered at $\boldsymbol{s}_j$, and $b(\boldsymbol{s}_j, \delta_j)$ denotes the disc centered at $\boldsymbol{s}_j$ with radius δ_j. Note that the above model-based reproduction number μ_j is event-specific since it depends on event marks through η_j, on the interaction ranges δ_j and τ_j, as well as on the event location $\boldsymbol{s}_j$ and time point t_j. If the model assumes unique interaction ranges δ and τ, a single reference number of secondary cases can be extrapolated from Equation 3 by imputing an unbounded domain $\boldsymbol{W} = \mathbb{R}^2$ and $T = \infty$ (Meyer *et al.* 2015).

Equation 3 can also be motivated by looking at a spatio-temporal version of the simple SIR model (1) wrapped into the `twinstim` class (2). This means: no endemic component, homogeneous force of infection ($\eta_j \equiv \beta$), homogeneous mixing in space ($f(x) = 1$, $\delta_j \equiv \infty$), and exponential decay of infectivity ($g(t) = e^{-\alpha t}$, $\tau_j \equiv \infty$). Then, for $T \to \infty$,

$$\mu = \beta \cdot \left[\int_0^{\infty} e^{-\alpha t}\, dt\right] \cdot \left[\int_{\boldsymbol{W}-\boldsymbol{s_j}} 1\, d\boldsymbol{s}\right] = \beta \cdot |\boldsymbol{W}|/\alpha ,$$

which corresponds to the basic reproduction number known from the simple SIR model by interpreting $|\boldsymbol{W}|$ as the population size, β as the transmission rate and α as the removal rate. If $\mu < 1$, the process is sub-critical, i.e., its eventual extinction is almost sure.

However, it is crucial to understand that in a full model with an endemic component, new infections may always occur via "immigration". Hence, reproduction numbers in `twinstim` are adjusted for infections occurring independently of previous infections. This also means that a misspecified endemic component may distort model-based reproduction numbers (Meyer *et al.* 2015). Furthermore, under-reporting and implemented control measures imply that the estimates are to be thought of as *effective* reproduction numbers.

Likelihood inference

The log-likelihood of the point process model (2) is a function of all parameters in the log-linear predictors $\nu_{[s][t]}$ and η_j and in the interaction functions f and g. It has the form

$$\left[\sum_{i=1}^{n} \log \lambda(\boldsymbol{s}_i, t_i)\right] - \int_0^T \int_{\boldsymbol{W}} \lambda(\boldsymbol{s}, t) \, \mathrm{d}\boldsymbol{s} \, \mathrm{d}t \, . \tag{5}$$

To estimate the model parameters, we maximize the above log-likelihood numerically using the quasi-Newton algorithm available through the R function `nlminb`. We thereby employ the analytical score function and an approximation of the expected Fisher information worked out by Meyer *et al.* (2012, Web Appendices A and B).

The space-time integral in the log-likelihood (5) poses no difficulties for the endemic component of $\lambda(\boldsymbol{s}, t)$, since $\nu_{[s][t]}$ is defined on a spatio-temporal grid. However, integration of the epidemic component involves two-dimensional integrals $\int_{\boldsymbol{R}_i} f(\|\boldsymbol{s}\|) \, \mathrm{d}\boldsymbol{s}$ over the influence regions $\boldsymbol{R}_i$, which are represented by polygons (as is $\boldsymbol{W}$). Similar integrals appear in the score function, where $f(\|\boldsymbol{s}\|)$ is replaced by partial derivatives with respect to kernel parameters. Calculation of these integrals is trivial for (piecewise) constant f, but otherwise requires numerical integration. For this purpose, the R package **polyCub** (Meyer 2015) offers cubature methods for polygonal domains as described in Meyer and Held (2014b, Section 2). For Gaussian f, we apply a midpoint rule with σ-adaptive bandwidth and use product Gauss cubature (Sommariva and Vianello 2007) to approximate the integrals in the score function. For the recently implemented power-law kernels, we established a cubature method which takes advantage of the assumed isotropy of spatial interaction such that numerical integration remains in only one dimension (Meyer and Held 2014b, Section 2.4). We also **memoise** (Wickham, Hester, and Müller 2016) the cubature function during log-likelihood maximization to avoid re-evaluations of the integrals with identical parameters of f.

Special cases: Single-component models

If the *epidemic* component is omitted in Equation 2, the point process model becomes equivalent to a Poisson regression model for aggregated counts. This provides a link to ecological regression approaches in general (Waller and Gotway 2004) and to the count data model `hhh4` illustrated in Section 5. To see this, recall that the endemic component $\nu_{[s][t]}$ is piecewise constant on the spatio-temporal grid with cells $([\boldsymbol{s}], [t])$. Hence the log-likelihood (5) of an endemic-only `twinstim` simplifies to a sum over all these cells,

$$\sum_{[\boldsymbol{s}],[t]} \left\{ Y_{[\boldsymbol{s}][t]} \log \nu_{[\boldsymbol{s}][t]} - |[\boldsymbol{s}]| \, |[t]| \, \nu_{[\boldsymbol{s}][t]} \right\} ,$$

where $Y_{[\boldsymbol{s}][t]}$ is the aggregated number of events observed in cell $([\boldsymbol{s}], [t])$, and $|[\boldsymbol{s}]|$ and $|[t]|$ denote cell area and length, respectively. Except for an additive constant, the above log-likelihood is equivalently obtained from the Poisson model $Y_{[\boldsymbol{s}][t]} \sim \mathrm{Po}(|[\boldsymbol{s}]| \, |[t]| \, \nu_{[\boldsymbol{s}][t]})$. This relation offers a means of code validation using the established `glm` function to fit an endemic-only `twinstim` model – see the examples in `help("glm_epidataCS")`.

If, in contrast, the *endemic* component is omitted, all events are necessarily triggered by other observed events. For such a model to be identifiable, a prehistory of events must exist to trigger the first event, and interaction typically needs to be unbounded such that each event can actually be linked to potential source events.

Extension: `twinstim` *with event types*

To model the example data on invasive meningococcal disease in the remainder of this section, we actually need to use an extended version $\lambda(\boldsymbol{s}, t, k)$ of Equation 2, which accounts for different event types k with own transmission dynamics. This introduces a further dimension in the point process, and the second log-likelihood component in Equation 5 accordingly splits into a sum over all event types. We refer to Meyer *et al.* (2012, Sections 2.4 and 3) for the technical details of this type-specific `twinstim` class. The basic idea is that the meningococcal finetypes share the same endemic pattern (e.g., seasonality), while infections of different finetypes are not associated via transmission. This means that the force of infection is restricted to previously infected individuals with the same bacterial finetype k, i.e., the epidemic sum in Equation 2 is over the set $I(\boldsymbol{s}, t, k) = I(\boldsymbol{s}, t) \cap \{j : k_j = k\}$. The implementation has limited support for type-dependent interaction functions f_{k_j} and g_{k_j} (not further considered here).

3.2. Data structure: `epidataCS`

The first step toward fitting a `twinstim` is to turn the relevant data into an object of the dedicated class `epidataCS`.[3] The primary ingredients of this class are a spatio-temporal point pattern (`events`) and its underlying observation region (`W`). An additional spatio-temporal grid (`stgrid`) holds (time-varying) areal-level covariates for the endemic regression part. We exemplify this data class by the `epidataCS` object for the 636 cases of invasive meningococcal disease in Germany originally analyzed by Meyer *et al.* (2012). It is already contained in the **surveillance** package as `data("imdepi")` and has been constructed as follows:

```
R> imdepi <- as.epidataCS(events = events, W = stateD, stgrid = stgrid,
+    qmatrix = diag(2), nCircle2Poly = 16)
```

The function `as.epidataCS` checks the consistency of the three data ingredients described in detail below. It also pre-computes auxiliary variables for model fitting, e.g., the individual influence regions (4), which are intersections of the observation region with discs approximated by polygons with `nCircle2Poly = 16` edges. The intersections are computed using functionality of the package **polyclip** (Johnson 2015). For multitype epidemics as in our example, the additional indicator matrix `qmatrix` specifies transmissibility across event types. An identity matrix corresponds to an independent spread of the event types, i.e., cases of one type can not produce cases of another type.

Data ingredients

The core `events` data must be provided in the form of a `SpatialPointsDataFrame` as defined by the package **sp** (Bivand *et al.* 2013):

```
R> summary(events)

Object of class SpatialPointsDataFrame
Coordinates:
   min  max
x 4039 4665
y 2710 3525
```

[3] The suffix "CS" indicates that the data-generating point process is indexed in continuous space.

```
Is projected: TRUE
proj4string :
[+init=epsg:3035 +units=km +proj=laea +lat_0=52 +lon_0=10 +x_0=4321000 +y_0=3210000
+ellps=GRS80 +no_defs]
Number of points: 636
Data attributes:
     time            tile    type       eps.t       eps.s        sex          agegrp
Min.   :   0.2  05354  : 34  B:336  Min.   :30  Min.   :200  female:292  [0,3)   :194
1st Qu.: 539.5  05370  : 27  C:300  1st Qu.:30  1st Qu.:200  male  :339  [3,19)  :279
Median :1155.0  11000  : 27         Median :30  Median :200  NA's  :  5  [19,Inf):162
Mean   :1192.7  05358  : 13         Mean   :30  Mean   :200              NA's    :  1
3rd Qu.:1808.0  05162  : 12         3rd Qu.:30  3rd Qu.:200
Max.   :2542.8  05382  : 12         Max.   :30  Max.   :200
                (Other):511
```

The associated event coordinates are residence postcode centroids, projected in the *European Terrestrial Reference System 1989* (in kilometer units) to enable Euclidean geometry. See the `spTransform`-methods in package **rgdal** (Bivand, Keitt, and Rowlingson 2015) for how to project latitude and longitude coordinates into a planar coordinate reference system (CRS). The data frame associated with these spatial coordinates ($\boldsymbol{s}_i$) contains a number of required variables and additional event marks (in the notation of Section 3.1: $\{(t_i, [\boldsymbol{s}_i], k_i, \tau_i, \delta_i, \boldsymbol{m}_i) : i = 1, \ldots, n\}$). For the IMD data, the event `time` is measured in days since the beginning of the observation period 2002–2008 and is subject to a tie-breaking procedure (described later). The `tile` column refers to the region of the spatio-temporal grid where the event occurred and here contains the official key of the administrative district of the patient's residence. There are two `type`s of events labeled as `"B"` and `"C"`, which refer to the serogroups of the two meningococcal finetypes *B:P1.7-2,4:F1-5* and *C:P1.5,2:F3-3* contained in the data. The `eps.t` and `eps.s` columns specify upper limits for temporal and spatial interaction, respectively. Here, the infectious period is assumed to last a maximum of 30 days and spatial interaction is limited to a 200 km radius for all cases. The latter has numerical advantages for a Gaussian interaction function f with a relatively small standard deviation. For a power-law kernel, however, this restriction will be dropped to enable occasional long-range transmission. The last two data attributes displayed in the above `event` summary are covariates from the case reports: the gender and age group of the patient.

For the observation region `W`, we use a polygon representation of Germany's boundary. Since the observation region defines the integration domain in the point process log-likelihood (5), the more detailed the polygons of `W` are the longer it will take to fit a `twinstim`. It is thus advisable to sacrifice some shape details for speed by reducing the polygon complexity, e.g., by applying one of the simplification methods available at MapShaper.org (Harrower and Bloch 2006). Alternative tools in R are **spatstat**'s `simplify.owin` procedure (Baddeley, Rubak, and Turner 2015) and the function `thinnedSpatialPoly` in package **maptools** (Bivand and Lewin-Koh 2016), which implements the Douglas and Peucker (1973) reduction method. The **surveillance** package already contains a simplified representation of Germany's boundaries:

```
R> load(system.file("shapes", "districtsD.RData", package = "surveillance"))
```

This file contains both the `SpatialPolygonsDataFrame districtsD` of Germany's 413 administrative districts as at January 1, 2009, as well as their union `stateD`. These boundaries are projected in the same CRS as the `events` data.

The `stgrid` input specific to the endemic model component is a simple data frame with (time-dependent) areal-level covariates, e.g., socio-economic or ecological characteristics. For our IMD example, we have:

```
      start stop  tile   area popdensity
1         0   31 01001   56.4     1557.1
2         0   31 01002  118.7     1996.6
3         0   31 01003  214.2      987.6
...     ...  ...   ...    ...        ...
34690  2526 2557 16075 1148.5       79.2
34691  2526 2557 16076  843.5      133.6
34692  2526 2557 16077  569.1      181.5
```

Numeric (`start`,`stop`] columns index the time periods and the factor variable `tile` identifies the regions of the grid. Note that the given time intervals (here: months) also define the resolution of possible time trends and seasonality of the piecewise constant endemic intensity. We choose monthly intervals to reduce package size and computational cost compared to the weekly resolution originally used by Meyer *et al.* (2012) and Meyer and Held (2014a). The above `stgrid` data frame thus consists of 7 (years) times 12 (months) blocks of 413 (districts) rows each. The `area` column gives the area of the respective `tile` in square kilometers (compatible with the CRS used for `events` and `W`). A geographic representation of the regions in `stgrid` is not required for model estimation, and is thus not part of the `epidataCS` class. In our example, the areal-level data only consists of the population density `popdensity`, whereas Meyer *et al.* (2012) additionally incorporated (lagged) weekly influenza counts by district as a time-dependent covariate.

Data handling and visualization

The generated `epidataCS` object `imdepi` is a simple list of the checked ingredients `events`, `stgrid`, `W` and `qmatrix`. Several methods for data handling and visualization are available for such objects as listed in Table 2 and briefly presented in the remainder of this section.

Printing an `epidataCS` object presents some metadata and the first 6 events by default:

```
R> imdepi
```

```
Observation period: 0 - 2557
Observation window (bounding box): [4031, 4672] x [2684, 3550]
Spatio-temporal grid (not shown): 84 time blocks x 413 tiles
Types of events: "B" "C"
Overall number of events: 636

   coordinates    time  tile type eps.t eps.s    sex   agegrp BLOCK start popdensity
1 (4112, 3203)  0.2117 05554    B    30   200   male   [3,19)     1     0      260.9
2 (4123, 3077)  0.7124 05382    C    30   200   male   [3,19)     1     0      519.4
3 (4412, 2916)  5.5910 09574    B    30   200 female [19,Inf)     1     0      209.4
4 (4203, 2880)  7.1170 08212    B    30   200 female   [3,19)     1     0     1665.6
5 (4128, 3223) 22.0595 05554    C    30   200   male   [3,19)     1     0      260.9
6 (4090, 3178) 24.9544 05170    C    30   200   male   [3,19)     1     0      454.7
[....]
```

During conversion to `epidataCS`, the last three columns `BLOCK` (time interval index), `start` and `popdensity` have been merged from the checked `stgrid` to the `events` data frame. The event marks including time and location can be extracted in a standard data frame by `marks(imdepi)`, and this is summarized by `summary(imdepi)`.

Display	Subset	Extract	Modify	Convert
`print`	`[`	`nobs`	`update`	`as.epidata`
`summary`	`head`	`marks`	`untie`	*`epidataCS2sts`*
`plot`	`tail`			
`animate`	`subset`			
`as.stepfun`				

Table 2: Generic and *non-generic* functions applicable to `epidataCS` objects.

A simple plot of the number of infectives as a function of time (Figure 1) can be obtained by the step function converter:

```
R> plot(as.stepfun(imdepi), xlim = summary(imdepi)$timeRange, xaxs = "i",
+    xlab = "Time [days]", ylab = "Current number of infectives", main = "")
```

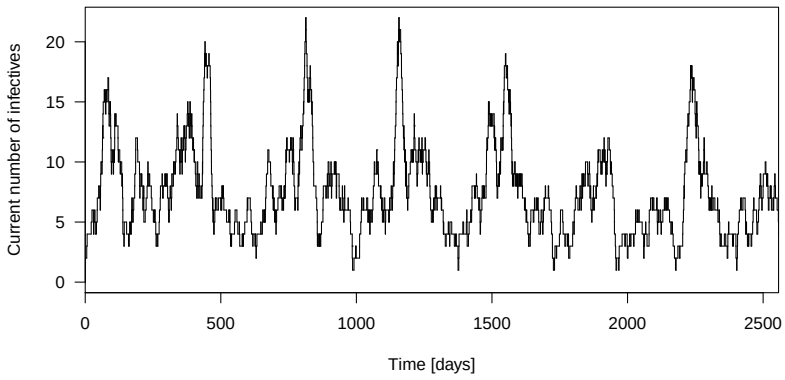
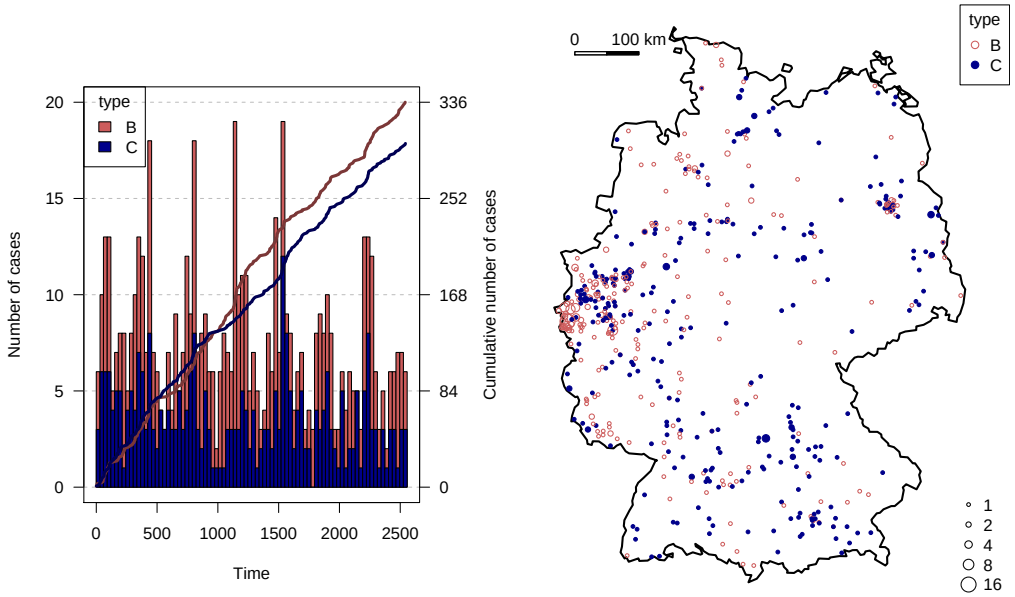


Figure 1: Time course of the number of infectives assuming infectious periods of 30 days.

The `plot`-method for `epidataCS` offers aggregation of the events over time or space:

```
R> plot(imdepi, "time", col = c("indianred", "darkblue"), ylim = c(0, 20))
R> plot(imdepi, "space", lwd = 2,
+    points.args = list(pch = c(1, 19), col = c("indianred", "darkblue")))
R> layout.scalebar(imdepi$W, scale = 100, labels = c("0", "100 km"), plot = TRUE)
```

The time-series plot (Figure 2a) shows the monthly aggregated number of cases by finetype in a stacked histogram as well as each type's cumulative number over time. The spatial plot (Figure 2b) shows the observation window `W` with the locations of all cases (by type), where the areas of the points are proportional to the number of cases at the respective location. Additional shading by the population is possible and exemplified in `help("plot.epidataCS")`.



(a) Temporal pattern.

(b) Spatial pattern.

Figure 2: Occurrence of the two finetypes viewed in the temporal and spatial dimensions.

The above static plots do not capture the space-time dynamics of epidemic spread. An animation may provide additional insight and can be produced by the corresponding `animate`-method. For instance, to look at the first year of the B-type in a weekly sequence of snapshots in a web browser (using facilities of the **animation** package of Xie 2013):

```
R> animation::saveHTML(
+    animate(subset(imdepi, type == "B"), interval = c(0, 365), time.spacing = 7),
+    nmax = Inf, interval = 0.2, loop = FALSE,
+    title = "Animation of the first year of type B events")
```

Selecting events from `epidataCS` as for the animation above is enabled by the `[`- and `subset`-methods, which return a new `epidataCS` object containing only the selected `events`.

A limited data sampling resolution may lead to tied event times or locations, which are in conflict with a continuous spatio-temporal point process model. For instance, a temporal residual analysis would suggest model deficiencies (Meyer *et al.* 2012, Figure 4), and a power-law kernel for spatial interaction may diverge if there are events with zero distance to potential source events (Meyer and Held 2014a). The function `untie` breaks ties by random shifts. This has already been applied to the event *times* in the provided `imdepi` data by subtracting a U(0,1)-distributed random number from the original dates. The event *coordinates* in the IMD data are subject to interval censoring at the level of Germany's postcode regions. A possible replacement for the given centroids would thus be a random location within the corresponding postcode area. Lacking a suitable shapefile, Meyer and Held (2014a) shifted all locations by a random vector with length up to half the observed minimum spatial separation:

```
R> eventDists <- dist(coordinates(imdepi$events))
R> (minsep <- min(eventDists[eventDists > 0]))

[1] 1.173

R> set.seed(321)
R> imdepi_untied <- untie(imdepi, amount = list(s = minsep / 2))
```

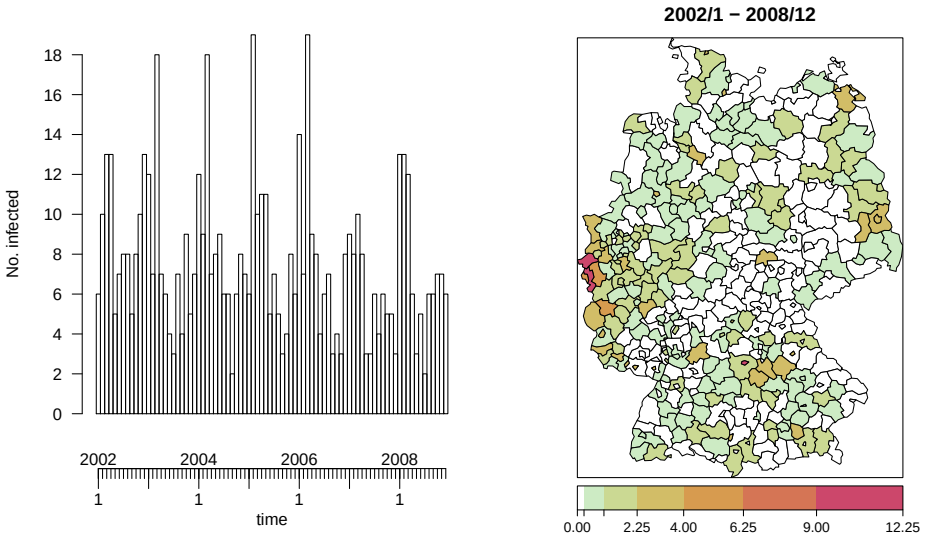
Note that random tie-breaking requires sensitivity analyses as discussed by Meyer and Held (2014a), but skipped here for the sake of brevity.

The `update`-method is useful to change the values of the maximum interaction ranges `eps.t` and `eps.s`, since it takes care of the necessary updates of the hidden auxiliary variables in an `epidataCS` object. For an unbounded interaction radius:

```
R> imdepi_untied_infeps <- update(imdepi_untied, eps.s = Inf)
```

Last but not least, `epidataCS` can be converted to the other classes `epidata` (Section 4) and `sts` (Section 5) by aggregation. The method `as.epidata.epidataCS` aggregates events by region (`tile`), and the function `epidataCS2sts` yields counts by region and time interval. The data could then, e.g., be analyzed by the multivariate time-series model presented in Section 5. We can also use visualization tools of the `sts` class, e.g., to produce Figure 3:

```
R> imdsts <- epidataCS2sts(imdepi, freq = 12, start = c(2002, 1), tiles = districtsD)
R> plot(imdsts, type = observed ~ time)
R> plot(imdsts, type = observed ~ unit, population = districtsD$POPULATION / 100000)
```



(a) Time series of monthly counts.

(b) Disease incidence (per 100 000 inhabitants).

Figure 3: IMD cases (joint types) aggregated as an `sts` object by month and district.

3.3. Modeling and inference

Having prepared the data as an object of class `epidataCS`, the function `twinstim` can be used to perform likelihood inference for conditional intensity models of the form (2). The main arguments for `twinstim` are the formulae of the `endemic` and `epidemic` linear predictors ($\nu_{[s][t]} = \exp(\texttt{endemic})$ and $\eta_j = \exp(\texttt{epidemic})$), and the spatial and temporal interaction functions `siaf` (f) and `tiaf` (g), respectively. Both formulae are parsed internally using the standard `model.frame` toolbox from package **stats** and thus can handle factor variables and interaction terms. While the `endemic` linear predictor incorporates time-dependent and/or areal-level covariates from `stgrid`, the `epidemic` formula may use both `stgrid` variables and event marks to be associated with the force of infection. For the interaction functions, several alternatives are predefined as listed in Table 3. They are applicable out-of-the-box and illustrated as part of the following modeling exercise for the IMD data. Own interaction functions can also be used provided their implementation obeys a certain structure, see `help("siaf")` and `help("tiaf")`, respectively.

Spatial (`siaf.*`)	Temporal (`tiaf.*`)
`constant`	`constant`
`gaussian`	`exponential`
`powerlaw`	`step`
`powerlawL`	
`step`	
`student`	

Table 3: Predefined spatial and temporal interaction functions.

Basic example

To illustrate statistical inference with `twinstim`, we will estimate several models for the simplified and "untied" IMD data presented in Section 3.2. In the endemic component, we include the district-specific population density as a multiplicative offset, a (centered) time trend, and a sinusoidal wave of frequency $2\pi/365$ to capture seasonality, where the `start` variable from `stgrid` measures time:

```
R> (endemic <- addSeason2formula(~offset(log(popdensity)) + I(start / 365 - 3.5),
+    period = 365, timevar = "start"))

~offset(log(popdensity)) + I(start/365 - 3.5) + sin(2 * pi *
    start/365) + cos(2 * pi * start/365)
```

See Held and Paul (2012, Section 2.2) for how such sine/cosine terms reflect seasonality. Because of the aforementioned integrations in the log-likelihood (5), it is advisable to first fit an endemic-only model to obtain reasonable start values for more complex epidemic models:

```
R> imdfit_endemic <- twinstim(endemic = endemic, epidemic = ~0,
+    data = imdepi_untied, subset = !is.na(agegrp))
```

We exclude the single case with unknown age group from this analysis since we will later estimate an effect of the age group on the force of infection.

Display	Extract	Modify	Other
print	nobs	update	simulate
summary	vcov	add1	*epitest*
xtable	coeflist	drop1	
plot	logLik	*stepComponent*	
intensityplot	extractAIC		
iafplot	profile		
checkResidualProcess	residuals		
	terms		
	R0		

Table 4: Generic and *non-generic* functions applicable to `twinstim` objects. Note that there is no need for specific `coef`, `confint`, `AIC` or `BIC` methods, since the respective default methods from package **stats** apply outright.

Many of the standard functions to access model fits in R are also implemented for `twinstim` fits (see Table 4). For example, we can produce the usual model summary:

```
R> summary(imdfit_endemic)

Call:
twinstim(endemic = endemic, epidemic = ~0, data = imdepi_untied,
    subset = !is.na(agegrp))

Coefficients of the endemic component:
                         Estimate Std. Error z value Pr(>|z|)
h.(Intercept)             -20.3683     0.0419 -486.24  < 2e-16 ***
h.I(start/365 - 3.5)       -0.0444     0.0200   -2.22    0.027 *
h.sin(2 * pi * start/365)   0.2733     0.0576    4.75  2.0e-06 ***
h.cos(2 * pi * start/365)   0.3509     0.0581    6.04  1.5e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

No epidemic component.

AIC:  19166
Log-likelihood: -9579
```

Because of the aforementioned equivalence of the endemic component with a Poisson regression model, the coefficients can be interpreted as log rate ratios in the usual way. For instance, the endemic rate is estimated to decrease by `1 - exp(coef(imdfit_endemic)[2])` $= 4.3\%$ per year. Coefficient correlations can be retrieved by the argument `correlation = TRUE` in the `summary` call just like for `summary.glm`, but may also be extracted via the standard `cov2cor(vcov(imdfit_endemic))`.

We now update the endemic model to take additional spatio-temporal dependence between events into account. Infectivity shall depend on the meningococcal finetype and the age group of the patient, and is assumed to be constant over time (default), $g(t) = \mathbb{1}_{(0,30]}(t)$, with a Gaussian distance-decay $f(x) = \exp\{-x^2/(2\sigma^2)\}$. This model was originally selected by Meyer *et al.* (2012) and can be fitted as follows:

```
R> imdfit_Gaussian <- update(imdfit_endemic, epidemic = ~type + agegrp,
+    siaf = siaf.gaussian(), start = c("e.(Intercept)" = -12.5, "e.siaf.1" = 2.75),
+    control.siaf = list(F = list(adapt = 0.25), Deriv = list(nGQ = 13)),
+    cores = 2 * (.Platform$OS.type == "unix"), model = TRUE)
```

To reduce the runtime of this example, we specified convenient `start` values for some parameters (others start at 0) and set `control.siaf` with a rather low number of nodes for the cubature of $f(\|\boldsymbol{s}\|)$ in the log-likelihood (via the midpoint rule) and $\frac{\partial f(\|\boldsymbol{s}\|)}{\partial \log \sigma}$ in the score function (via product Gauss cubature). On Unix-alikes, these numerical integrations can be performed in parallel using the "multicore" functions `mclapply` *et al.* from the base package **parallel**, here with `cores = 2` processes. For later generation of an `intensityplot`, the `model` environment is retained.

	RR	95% CI	p-value
`h.I(start/365 - 3.5)`	0.955	0.91–1.00	0.039
`h.sin(2 * pi * start/365)`	1.243	1.09–1.41	0.0008
`h.cos(2 * pi * start/365)`	1.375	1.21–1.56	<0.0001
`e.typeC`	0.402	0.24–0.68	0.0007
`e.agegrp[3,19)`	2.000	1.06–3.78	0.033
`e.agegrp[19,Inf)`	0.776	0.32–1.91	0.58

Table 5: Estimated rate ratios (RR) and associated Wald confidence intervals (CI) for endemic (`h.`) and epidemic (`e.`) terms. This table was generated by `xtable(imdfit_Gaussian)`.

Table 5 shows the output of `twinstim`'s `xtable` method (Dahl 2016), which provides rate ratios for the endemic and epidemic effects. The alternative `toLatex` method simply translates the `summary` table of coefficients to LaTeX without `exp`-transformation. On the subject-matter level, we can conclude from Table 5 that the meningococcal finetype of serogroup C is less than half as infectious as the B-type, and that patients in the age group 3 to 18 years are estimated to cause twice as many secondary infections as infants aged 0 to 2 years.

Model-based effective reproduction numbers

The event-specific reproduction numbers (3) can be extracted from fitted `twinstim` objects via the `R0` method. For the above IMD model, we obtain the following mean numbers of secondary infections by finetype:

```
R> R0_events <- R0(imdfit_Gaussian)
R> tapply(R0_events, marks(imdepi_untied)[names(R0_events), "type"], mean)

      B       C
0.21614 0.09576
```

Confidence intervals can be obtained via Monte Carlo simulation, where Equation 3 is repeatedly evaluated with parameters sampled from the asymptotic multivariate normal distribution of the maximum likelihood estimate. For this purpose, the `R0`-method takes an argument `newcoef`, which is exemplified in `help("R0")`.

Interaction functions

Figure 4 shows several estimated spatial interaction functions, which can be plotted by, e.g., `plot(imdfit_Gaussian, which = "siaf")`. Meyer and Held (2014a) found that a power-law decay of spatial interaction is more appropriate than a Gaussian kernel to describe the spread of human infectious diseases. The power-law kernel concentrates on short-range interaction, but also exhibits a heavier tail reflecting occasional transmission over large distances. To use the power-law kernel $f(x) = (x + \sigma)^{-d}$, we switch to the prepared `epidataCS` object with `eps.s = Inf` and update the previous Gaussian model as follows:

```
R> imdfit_powerlaw <- update(imdfit_Gaussian, data = imdepi_untied_infeps,
+    siaf = siaf.powerlaw(), control.siaf = NULL,
+    start = c("e.(Intercept)" = -6.2, "e.siaf.1" = 1.5, "e.siaf.2" = 0.9))
```

Table 3 also lists the step function kernel as an alternative, which is particularly useful for two reasons. First, it is a more flexible approach since it estimates interaction between the given knots without assuming an overall functional form. Second, the spatial integrals in the log-likelihood can be computed analytically for the step function kernel, which therefore offers a quick estimate of spatial interaction. We update the Gaussian model to use four steps at log-equidistant knots up to an interaction range of 100 km:

```
R> imdfit_step4 <- update(imdfit_Gaussian, data = imdepi_untied_infeps,
+    siaf = siaf.step(exp(1:4 * log(100) / 5), maxRange = 100), control.siaf = NULL,
+    start = c("e.(Intercept)" = -10, setNames(-2:-5, paste0("e.siaf.", 1:4))))
```

Figure 4 suggests that the estimated step function is in line with the power law.

For the temporal interaction function $g(t)$, model updates and plots are similarly possible, e.g., `update(imdfit_Gaussian, tiaf = tiaf.exponential())`. However, the events in the IMD data are too rare to infer the time-course of infectivity with confidence.

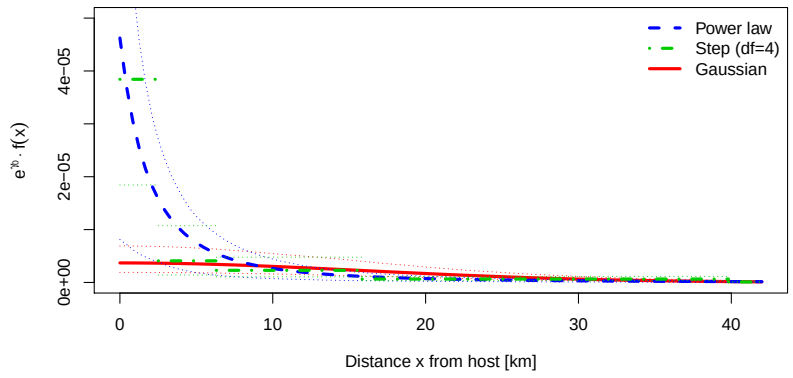


Figure 4: Various estimates of spatial interaction (scaled by the epidemic intercept γ_0). The standard deviation of the Gaussian kernel is estimated to be $\hat{\sigma} = 16.00$ (95% CI: 13.65–18.75), and the estimated power-law parameters are $\hat{\sigma} = 4.64$ (95% CI: 1.82–11.84) and $\hat{d} = 2.49$ (95% CI: 1.81–3.42).

Model selection

```
R> AIC(imdfit_endemic, imdfit_Gaussian, imdfit_powerlaw, imdfit_step4)

                df   AIC
imdfit_endemic   4 19166
imdfit_Gaussian  9 18967
imdfit_powerlaw 10 18940
imdfit_step4    12 18933
```

Akaike's Information Criterion (AIC) suggests superiority of the power-law vs. the Gaussian model and the endemic-only model. The more flexible step function yields the best AIC value but its shape strongly depends on the chosen knots and is not guaranteed to be monotonically decreasing. The function `stepComponent` – a wrapper around the `step` function from **stats** – can be used to perform AIC-based stepwise selection within a given model component.

Model diagnostics

Two other plots are implemented for `twinstim` objects. Figure 5 shows an `intensityplot` of the fitted "ground" intensity $\sum_{k=1}^{2} \int_{\boldsymbol{W}} \hat{\lambda}(\boldsymbol{s}, t, k)\, \mathrm{d}\boldsymbol{s}$ aggregated over both event types:

```
R> intensityplot(imdfit_powerlaw, which = "total", aggregate = "time", types = 1:2)
```

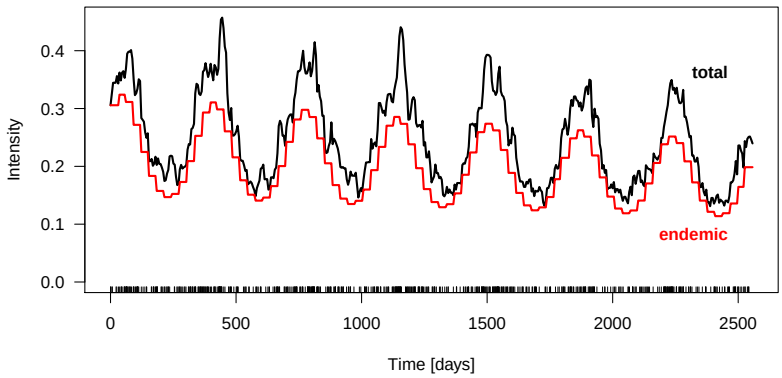
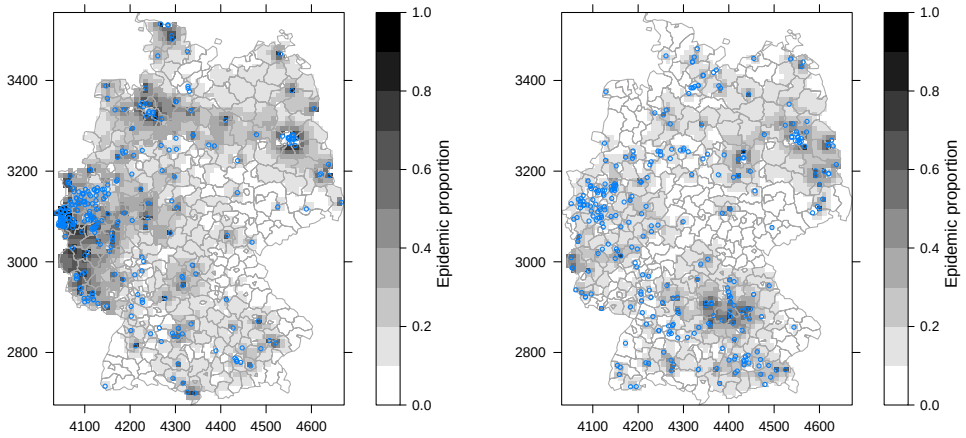


Figure 5: Fitted "ground" intensity process aggregated over space and both types.

The estimated endemic intensity component has also been added to the plot. It exhibits strong seasonality and a slow negative trend. The proportion of the endemic intensity is rather constant along time since no major outbreaks occurred. This proportion can be visualized separately by specifying `which = "endemic proportion"` in the above call.

Spatial `intensityplots` can be produced via `aggregate = "space"` and require a geographic representation of `stgrid`. Figure 6 shows the accummulated epidemic proportion by event type. It is naturally high in regions with a large number of cases and even more so if the population density is low. The function `epitest` offers a model-based global test for epidemicity, while `knox` and `stKtest` implement related classical approaches (Meyer *et al.* 2015).



(a) Type B. (b) Type C.

Figure 6: Epidemic proportion of the fitted intensity process accumulated over time by type.

Another diagnostic tool is the function `checkResidualProcess`, which transforms the temporal "residual process" in such a way that it exhibits a uniform distribution and lacks serial correlation if the fitted model describes the true CIF well (see Ogata 1988, Section 3.3). These properties can be checked graphically as in Figure 7 produced by:

```
R> checkResidualProcess(imdfit_powerlaw)
```

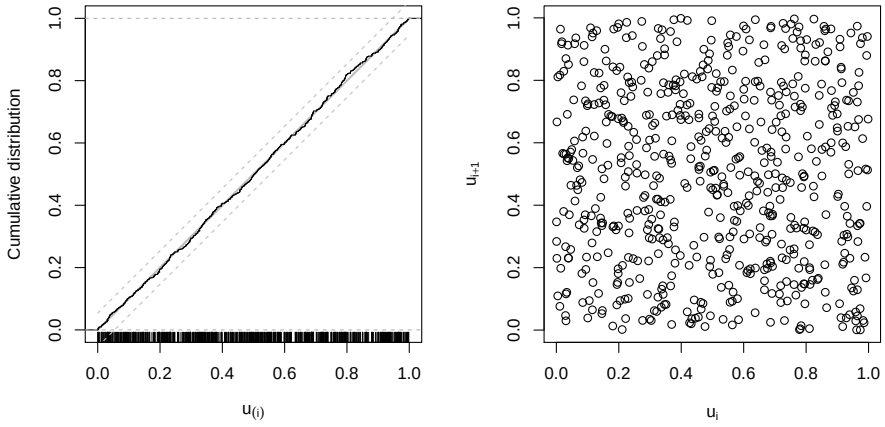


Figure 7: The left plot shows the `ecdf` of the transformed residuals with a 95% confidence band obtained by inverting the corresponding Kolmogorov-Smirnov test (no evidence for deviation from uniformity). The right-hand plot suggests absence of serial correlation.

3.4. Simulation

To identify regions with unexpected IMD dynamics, Meyer *et al.* (2012) compared the observed numbers of cases by district to the respective 2.5% and 97.5% quantiles of 100 simulations from the selected model. Furthermore, simulations allow us to investigate the stochastic volatility of the endemic-epidemic process, to obtain probabilistic forecasts, and to perform parametric bootstrap of the spatio-temporal point pattern.

The simulation algorithm we apply is described in Meyer *et al.* (2012, Section 4). It requires a geographic representation of the `stgrid`, as well as functionality for sampling locations from the spatial kernel $f_2(\boldsymbol{s}) := f(\|\boldsymbol{s}\|)$. This is implemented for all predefined spatial interaction functions listed in Table 3. Event marks are by default sampled from their respective empirical distribution in the original data. The following code runs 30 simulations over the last two years based on the estimated power-law model:

```
R> imdsims <- simulate(imdfit_powerlaw, nsim = 30, seed = 1, t0 = 1826, T = 2555,
+    data = imdepi_untied_infeps, tiles = districtsD)
```

Figure 8 shows the cumulative number of cases from the simulations appended to the first five years of data. Extracting a single simulation (e.g., `imdsims[[1]]`) yields an object of the class `simEpidataCS`, which extends `epidataCS`. It carries additional components from the generating model to enable an `R0`-method and `intensityplot`s for simulated data. A special feature of such simulations is that the source of each event is actually known:

```
R> table(imdsims[[1]]$events$source > 0, exclude = NULL)

FALSE  TRUE  <NA>
  112    25     8
```

The stored `source` value is 0 for endemic events, `NA` for events of the prehistory but still infective at `t0`, and otherwise corresponds to the row index of the infective source. Averaged over all 30 simulations, the proportion of events triggered by previous events is 0.2179.

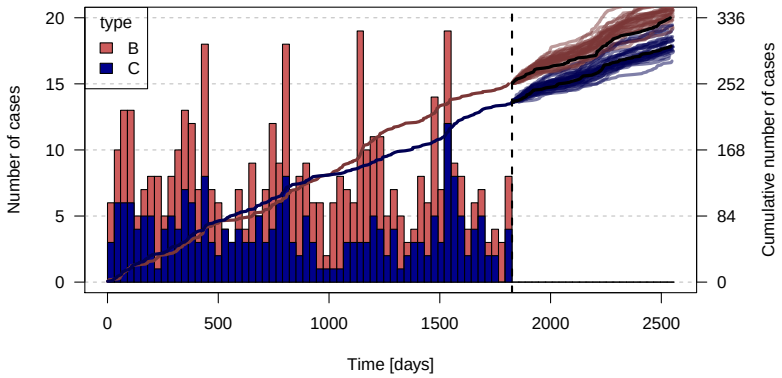


Figure 8: Simulation-based forecast of the cumulative number of cases by finetype in the last two years. The black lines correspond to the observed numbers.

4. SIR event history of a fixed population

The endemic-epidemic multivariate point process model "`twinSIR`" is designed for individual-level surveillance data of a fixed population of which the complete SIR event history is assumed to be known. As an illustrative example, we use a particularly well-documented measles outbreak among children of the isolated German village Hagelloch in the year 1861, which has previously been analyzed by, e.g., Neal and Roberts (2004). Other potential applications include farm-level data as well as epidemics across networks. We start by describing the general model class in Section 4.1. Section 4.2 introduces the example data and the associated class `epidata`, and Section 4.3 presents the core functionality of fitting and analyzing such data using `twinSIR`. Due to the many similarities with the `twinstim` framework covered in Section 3, we condense the `twinSIR` treatment accordingly.

4.1. Model class: `twinSIR`

The point process model `twinstim` (Section 3) is indexed in a continuous spatial domain, i.e., the set of possible event locations consists of the whole observation region and is thus infinite. However, if infections can only occur at a known discrete set of sites, such as for livestock diseases among farms, the conditional intensity function formally becomes $\lambda_i(t)$. It characterizes the instantaneous rate of infection of individual i at time t, given the sets $S(t)$ and $I(t)$ of susceptible and infectious individuals, respectively (just before time t). In a similar regression view as in Section 3, Höhle (2009) proposed the endemic-epidemic multivariate temporal point process model "`twinSIR`":

$$\lambda_i(t) = \lambda_0(t)\,\nu_i(t) + \sum_{j \in I(t)} \left\{ f(d_{ij}) + \boldsymbol{w}_{ij}^\top \boldsymbol{\alpha}^{(w)} \right\} , \tag{6}$$

if $i \in S(t)$, i.e., if individual i is currently susceptible, and $\lambda_i(t) = 0$ otherwise. The rate decomposes into two components. The endemic component consists of a Cox proportional hazards formulation containing a semi-parametric baseline hazard $\lambda_0(t)$ and a log-linear predictor $\nu_i(t) = \exp\left(\boldsymbol{z}_i(t)^\top \boldsymbol{\beta}\right)$ of covariates modeling infection from external sources. Furthermore, an additive epidemic component captures transmission from the set $I(t)$ of currently infectious individuals. The force of infection of individual i depends on the distance d_{ij} to each infective source $j \in I(t)$ through a distance kernel

$$f(u) = \sum_{m=1}^{M} \alpha_m^{(f)} B_m(u) \;\geq 0 \,, \tag{7}$$

which is represented by a linear combination of non-negative basis functions B_m with the $\alpha_m^{(f)}$'s being the respective coefficients. For instance, f could be modelled by a B-spline (Fahrmeir, Kneib, Lang, and Marx 2013, Section 8.1), and d_{ij} could refer to the Euclidean distance $\|\boldsymbol{s}_i - \boldsymbol{s}_j\|$ between the individuals' locations $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$, or to the geodesic distance between the nodes i and j in a network. The distance-based force of infection is modified additively by a linear predictor of covariates $\boldsymbol{w}_{ij}$ describing the interaction of individuals i and j further. Hence, the whole epidemic component of Equation 6 can be written as a single linear predictor $\boldsymbol{x}_i(t)^\top \boldsymbol{\alpha}$ by interchanging the summation order to

$$\sum_{m=1}^{M} \alpha_m^{(f)} \sum_{j \in I(t)} B_m(d_{ij}) + \sum_{k=1}^{K} \alpha_k^{(w)} \sum_{j \in I(t)} w_{ijk} = \boldsymbol{x}_i(t)^\top \boldsymbol{\alpha} \,, \tag{8}$$

such that $\boldsymbol{x}_i(t)$ comprises all epidemic terms summed over $j \in I(t)$. Note that the use of additive covariates $\boldsymbol{w}_{ij}$ on top of the distance kernel in (6) is different from `twinstim`'s multiplicative approach in (2). One advantage of the additive approach is that the subsequent linear decomposition of the distance kernel allows one to gather all parts of the epidemic component in a single linear predictor. Hence, the above model represents a CIF extension of what in the context of survival analysis is known as an additive-multiplicative hazard model (Martinussen and Scheike 2002). As a consequence, the `twinSIR` model could in principle be fitted with the **timereg** package (Scheike and Martinussen 2006), which yields estimates for the cumulative hazards. However, Höhle (2009) chooses a more direct inferential approach: To ensure that the CIF $\lambda_i(t)$ is non-negative, all covariates are encoded such that the components of $\boldsymbol{w}_{ij}$ are non-negative. Additionally, the parameter vector $\boldsymbol{\alpha}$ is constrained to be non-negative. Subsequent parameter inference is then based on the resulting constrained penalized likelihood which gives directly interpretable estimates of $\boldsymbol{\alpha}$. Future work could investigate the potential of a multiplicative approach for the epidemic component in `twinSIR`.

4.2. Data structure: `epidata`

New SIR-type event data typically arrive in the form of a simple data frame with one row per individual and sequential event time points as columns. For the 1861 Hagelloch measles epidemic, such a data set of the 188 affected children is contained in the **surveillance** package:

```
R> data("hagelloch")
R> head(hagelloch.df, n = 5)
```

```
  PN    NAME FN HN AGE    SEX        PRO        ERU        CL DEAD IFTO SI
1  1 Mueller 41 61   7 female 1861-11-21 1861-11-25 1st class <NA>   45 10
2  2 Mueller 41 61   6 female 1861-11-23 1861-11-27 1st class <NA>   45 12
3  3 Mueller 41 61   4 female 1861-11-28 1861-12-02 preschool <NA>  172  9
4  4 Seibold 61 62  13   male 1861-11-27 1861-11-28 2nd class <NA>  180 10
5  5  Motzer 42 63   8 female 1861-11-22 1861-11-27 1st class <NA>   45 11
               C PR CA NI GE TD   TM x.loc y.loc  tPRO  tERU tDEAD    tR    tI
1 no complicatons  4  4  3  1 NA   NA 142.5 100.0 22.71 26.23    NA 29.23 21.71
2 no complicatons  4  4  3  1  3 40.3 142.5 100.0 24.21 28.79    NA 31.79 23.21
3 no complicatons  4  4  3  2  1 40.5 142.5 100.0 29.59 33.69    NA 36.69 28.59
4 no complicatons  1  1  1  1  3 40.7 165.0 102.5 28.12 29.03    NA 32.03 27.12
5 no complicatons  5  3  2  1 NA   NA 145.0 120.0 23.06 28.42    NA 31.42 22.06
```

The `help("hagelloch")` contains a description of all columns. Here we concentrate on the event columns `PRO` (appearance of prodromes), `ERU` (eruption), and `DEAD` (day of death if during the outbreak). We take the day on which the index case developed first symptoms, 30 October 1861 (`min(hagelloch.df$PRO)`), as the start of the epidemic, i.e., we condition on this case being initially infectious. As for `twinstim`, the property of point processes that concurrent events have zero probability requires special treatment. Ties are due to the interval censoring of the data to a daily basis – we broke these ties by adding random jitter to the event times within the given days. The resulting columns `tPRO`, `tERU`, and `tDEAD` are relative to the defined start time. Following Neal and Roberts (2004), we assume that each child becomes infectious (S $\rightarrow$ I event at time `tI`) one day before the appearance of prodromes, and is removed from the epidemic (I $\rightarrow$ R event at time `tR`) three days after the appearance of rash or at the time of death, whichever comes first.

For further processing of the data, we convert `hagelloch.df` to the standardized `epidata` structure for `twinSIR`. This is done by the converter function `as.epidata`, which also checks consistency and optionally pre-calculates the epidemic terms $\boldsymbol{x}_i(t)$ of Equation 8 to be incorporated in a `twinSIR` model. The following call generates the `epidata` object `hagelloch`:

```
R> hagelloch <- as.epidata(hagelloch.df,
+    t0 = 0, tI.col = "tI", tR.col = "tR",
+    id.col = "PN", coords.cols = c("x.loc", "y.loc"),
+    f = list(household    = function(u) u == 0,
+             nothousehold = function(u) u > 0),
+    w = list(c1 = function (CL.i, CL.j) CL.i == "1st class" & CL.j == CL.i,
+             c2 = function (CL.i, CL.j) CL.i == "2nd class" & CL.j == CL.i),
+    keep.cols = c("SEX", "AGE", "CL"))
```

The coordinates (`x.loc`, `y.loc`) correspond to the location of the household the child lives in and are measured in meters. Note that `twinSIR` allows for tied locations of individuals, but assumes the relevant spatial location to be fixed during the entire observation period. By default, the Euclidean distance between the given coordinates will be used. Alternatively, `as.epidata` also accepts a pre-computed distance matrix via its argument `D` without requiring spatial coordinates. The argument `f` lists distance-dependent basis functions B_m for which the epidemic terms $\sum_{j \in I(t)} B_m(d_{ij})$ shall be generated. Here, `household` ($x_{i,H}(t)$) and `nothousehold` ($x_{i,\bar{H}}(t)$) count for each child the number of currently infective children in its household and outside its household, respectively. Similar to Neal and Roberts (2004), we also calculate the covariate-based epidemic terms `c1` ($x_{i,c1}(t)$) and `c2` ($x_{i,c2}(t)$) counting the number of currently infective classmates. Note from the corresponding definitions of w_{ij1} and w_{ij2} in `w` that `c1` is always zero for children of the second class and `c2` is always zero for children of the first class. For pre-school children, both variables equal zero over the whole period. By the last argument `keep.cols`, we choose to only keep the covariates `SEX`, `AGE`, and school `CL`ass from `hagelloch.df`.

The first few rows of the generated `epidata` object are shown below:

```
R> head(hagelloch, n = 5)
```

```
  BLOCK id start  stop atRiskY event Revent x.loc y.loc    SEX AGE        CL
1     1  1     0 1.136       1     0      0 142.5 100.0 female   7 1st class
2     1  2     0 1.136       1     0      0 142.5 100.0 female   6 1st class
3     1  3     0 1.136       1     0      0 142.5 100.0 female   4 preschool
4     1  4     0 1.136       1     0      0 165.0 102.5   male  13 2nd class
5     1  5     0 1.136       1     0      0 145.0 120.0 female   8 1st class
  household nothousehold c1 c2
1         0            1  0  0
2         0            1  0  0
3         0            1  0  0
4         0            1  0  1
5         0            1  0  0
```

The `epidata` structure inherits from counting processes as implemented by the `Surv` class of package **survival** (Therneau 2015) and also used in **timereg** (Scheike and Martinussen 2006). Specifically, the observation period is splitted up into consecutive time intervals (`start`; `stop`]

of constant conditional intensities. As the CIF $\lambda_i(t)$ of Equation (6) only changes at time points, where the set of infectious individuals $I(t)$ or some endemic covariate in $\nu_i(t)$ change, those occurrences define the break points of the time intervals. Altogether, the `hagelloch` event history consists of 375 time `BLOCK`s of 188 rows, where each row describes the state of individual `id` during the corresponding time interval. The susceptibility status and the I- and R-events are captured by the columns `atRiskY`, `event` and `Revent`, respectively. The `atRiskY` column indicates if the individual is at risk of becoming infected in the current interval. The event columns indicate, which individual was infected or removed at the `stop` time. Note that at most one entry in the `event` and `Revent` columns is 1, all others are 0.

Apart from being the input format for `twinSIR` models, the `epidata` class has several associated methods (Table 6), which are similar in spirit to the methods described for `epidataCS`.

Display	Subset	Modify
`print`	`[`	`update`
`summary`		
`plot`		
`animate`		
`stateplot`		

Table 6: Generic and *non-generic* functions applicable to `epidata` objects.

For example, Figure 9 illustrates the course of the Hagelloch measles epidemic by counting processes for the number of susceptible, infectious and removed children, respectively. Figure 10 shows the locations of the households. An `animate`d map can also be produced to view the households' states over time and a `stateplot` shows the changes for a selected unit.

```
R> plot(hagelloch, xlab = "Time [days]")
```

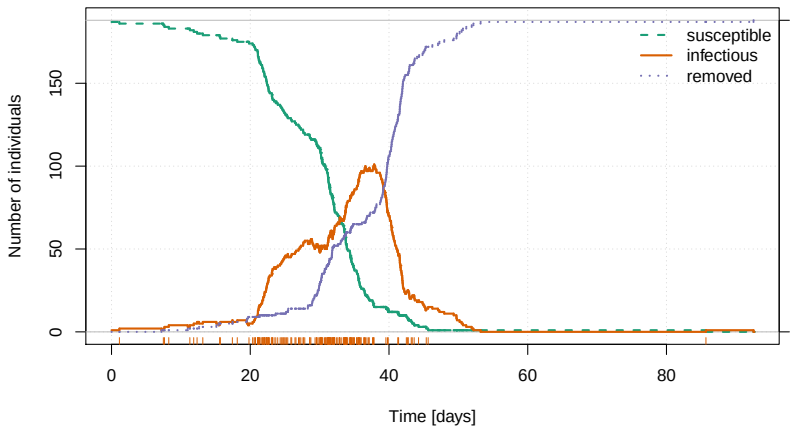


Figure 9: Evolution of the 1861 Hagelloch measles epidemic in terms of the numbers of susceptible, infectious, and recovered children. The bottom `rug` marks the infection times `tI`.

```
R> hagelloch_coords <- summary(hagelloch)$coordinates
R> plot(hagelloch_coords, xlab = "x [m]", ylab = "y [m]",
+    pch = 15, asp = 1, cex = sqrt(multiplicity(hagelloch_coords)))
R> legend(x = "topleft", pch = 15, legend = c(1, 4, 8), pt.cex = sqrt(c(1, 4, 8)),
+    title = "Household size")
```

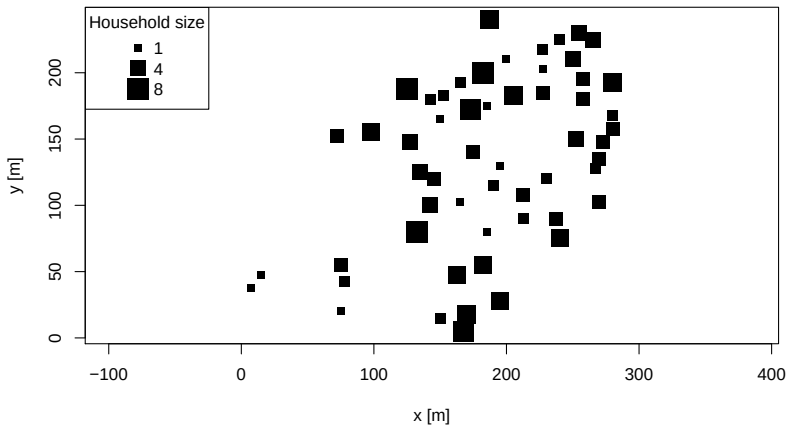


Figure 10: Spatial locations of the Hagelloch households. The size of each dot is proportional to the number of children in the household.

4.3. Modeling and inference

Basic example

To illustrate the flexibility of `twinSIR` we will analyze the Hagelloch data using class room and household indicators similar to Neal and Roberts (2004). We include an additional endemic background rate $\exp(\beta_0)$, which allows for multiple outbreaks triggered by external sources. Consequently, we do not need to ignore the child that got infected about one month after the end of the main epidemic (see the last event mark in Figure 9), as, e.g., done in a thorough network-based analysis of the Hagelloch data by Groendyke, Welch, and Hunter (2012). Altogether, the CIF for a child i is modeled as

$$\lambda_i(t) = Y_i(t) \cdot \left[\exp(\beta_0) + \alpha_H x_{i,H}(t) + \alpha_{c1} x_{i,c1}(t) + \alpha_{c2} x_{i,c2}(t) + \alpha_{\bar{H}} x_{i,\bar{H}}(t)\right] , \qquad (9)$$

where $Y_i(t) = \mathbb{1}(i \in S(t))$ is the at-risk indicator. By counting the number of infectious classmates separately for both school classes as described in the previous section, we allow for class-specific effects α_{c1} and α_{c2} on the force of infection. The model is estimated by maximum likelihood (Höhle 2009) using the following call:

```
R> hagellochFit <- twinSIR(~household + c1 + c2 + nothousehold, data = hagelloch)
R> summary(hagellochFit)
```

```
Call:
twinSIR(formula = ~household + c1 + c2 + nothousehold, data = hagelloch)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
household         0.026868   0.006113    4.39  1.1e-05 ***
c1                0.023892   0.005026    4.75  2.0e-06 ***
c2                0.002932   0.000755    3.88   0.0001 ***
nothousehold      0.000831   0.000142    5.87  4.3e-09 ***
cox(logbaseline) -7.362644   0.887989   -8.29  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total number of infections:  187

One-sided AIC: 1245        (simulated penalty weights)
Log-likelihood: -619
Number of log-likelihood evaluations: 119
```

The results show, e.g., a $0.0239 / 0.0029 = 8.149$ times higher transmission between individuals in the 1st class than in the 2nd class. Furthermore, an infectious housemate adds $0.0269 / 0.0008 = 32.32$ times as much infection pressure as infectious children outside the household. The endemic background rate of infection in a population with no current measles cases is estimated to be $\exp(\hat{\beta}_0) = \exp(-7.363) = 0.0006345$. An associated Wald confidence interval (CI) based on the asymptotic normality of the maximum likelihood estimator (MLE) can be obtained by `exp`-transforming the `confint` for β_0:

```
R> exp(confint(hagellochFit, parm = "cox(logbaseline)"))

                     2.5 %   97.5 %
cox(logbaseline) 0.0001113 0.003617
```

Note that Wald confidence intervals for the epidemic parameters $\boldsymbol{\alpha}$ are to be treated carefully, because their construction does not take the restricted parameter space into account. For more adequate statistical inference, the behavior of the log-likelihood near the MLE can be investigated using the `profile`-method for `twinSIR` objects. For instance, to evaluate the normalized profile log-likelihood of α_{c1} and α_{c2} on an equidistant grid of 25 points within the corresponding 95% Wald CIs, we do:

```
R> prof <- profile(hagellochFit,
+    list(c(match("c1", names(coef(hagellochFit))), NA, NA, 25),
+         c(match("c2", names(coef(hagellochFit))), NA, NA, 25)))
```

The profiling result contains 95% highest likelihood based CIs for the parameters, as well as the Wald CIs for comparison:

```
R> prof$ci.hl

   idx   hl.low    hl.up wald.low  wald.up      mle
c1   2 0.015219 0.034969 0.014041 0.033744 0.023892
c2   3 0.001576 0.004535 0.001453 0.004411 0.002932
```

The entire functional form of the normalized profile log-likelihood on the requested grid as stored in `prof$lp` can be visualized by:

```
R> plot(prof)
```

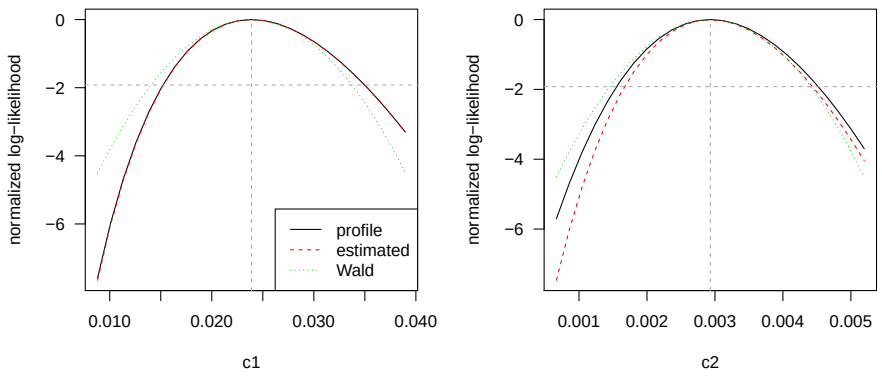


Figure 11: Normalized log-likelihood for α_{c1} and α_{c2} when fitting the `twinSIR` model formulated in Equation (9) to the Hagelloch data.

Model diagnostics

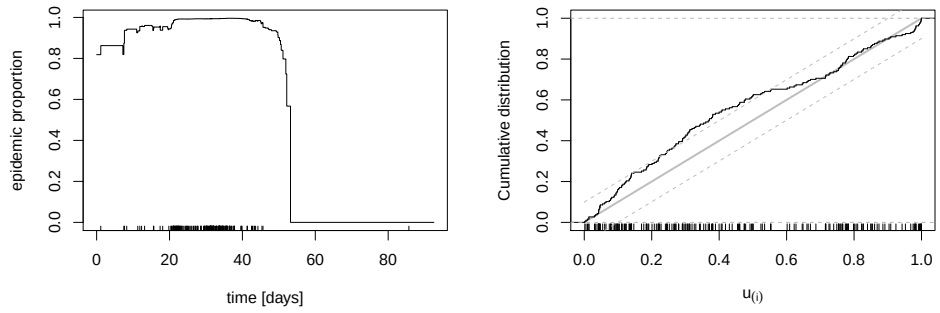
Display	Extract	Other
`print`	`vcov`	`simulate`
`summary`	`logLik`	
`plot`	`AIC`	
`intensityplot`	`extractAIC`	
`checkResidualProcess`	`profile`	
	`residuals`	

Table 7: Generic and *non-generic* functions for `twinSIR`. There are no specific `coef` or `confint` methods, since the respective default methods from package **stats** apply outright.

Table 7 lists all methods for the `twinSIR` class. For example, to investigate how the CIF decomposes into endemic and epidemic intensity over time, we produce Figure 12a by:

```
R> plot(hagellochFit, which = "epidemic proportion", xlab = "time [days]")
```

Note that the last infection was necessarily caused by the endemic component since there were no more infectious children in the observed population which could have triggered the new case. We can also inspect temporal Cox-Snell-like `residuals` of the fitted point process using the function `checkResidualProcess` as for the spatio-temporal point process models in Section 3.3. The resulting Figure 12b reveals some deficiencies of the model in describing the waiting times between events, which might be related to the assumption of fixed infection periods.



(a) Epidemic proportion.

(b) Transformed residuals.

Figure 12: Diagnostic plots for the `twinSIR` model formulated in Equation 9.

Finally, `twinSIR`'s `AIC`-method computes the one-sided AIC (Hughes and King 2003) as described in Höhle (2009), which can be used for model selection under positivity constraints on $\boldsymbol{\alpha}$. For instance, we may consider a more flexible model for local spread using a step function for the distance kernel $f(u)$ in Equation 7. An updated model with $B_1 = I_{(0;100)}(u)$, $B_2 = I_{[100;200)}(u)$, $B_3 = I_{[200;\infty)}(u)$ can be fitted as follows:

```
R> knots <- c(100, 200)
R> fstep <- list(
+    B1 = function(D) D > 0 & D < knots[1],
+    B2 = function(D) D >= knots[1] & D < knots[2],
+    B3 = function(D) D >= knots[2])
R> hagellochFit_fstep <- twinSIR(
+    ~household + c1 + c2 + B1 + B2 + B3,
+    data = update(hagelloch, f = fstep))

R> set.seed(1)
R> AIC(hagellochFit, hagellochFit_fstep)

                   df  AIC
hagellochFit        5 1245
hagellochFit_fstep  7 1246
```

Hence the simpler model with just a `nothousehold` component instead of the more flexible distance-based step function is preferred. A random seed was set since the parameter penalty in the one-sided AIC is determined by Monte Carlo simulation. The algorithm is described in Silvapulle and Sen (2005, p. 79, Simulation 3) and involves quadratic programming using package **quadprog** (Turlach 2013).

4.4. Simulation

Simulation from fitted `twinSIR` models is described in detail in Höhle (2009, Section 4). The implementation is made available by an appropriate `simulate`-method for class `twinSIR`. Because both the algorithm and the call are similar to the invocation on `twinstim` objects (Section 3.4), we skip the illustration here and refer to `help("simulate.twinSIR")`.

5. Areal time series of counts

In public health surveillance, routine reports of infections to public health authorities give rise to spatio-temporal data, which are usually made available in the form of aggregated counts by region and period. The Robert Koch Institute (RKI) in Germany, for example, maintains a database of cases of notifiable diseases, which can be queried via the *SurvStat@RKI*[4] online service. As an illustrative example, we use weekly counts of measles infections by district in the Weser-Ems region of Lower Saxony, Germany, 2001–2002. These spatio-temporal count data constitute the response Y_{it}, $i = 1, \ldots, 17$ (districts), $t = 1, \ldots, 104$ (weeks), for our illustration of the endemic-epidemic multivariate time-series model "`hhh4`". We start by describing the general model class in Section 5.1. Section 5.2 introduces the data and the associated `S4`-class `sts` ("surveillance time series"). In Section 5.3, a simple model for the measles data based on the original analysis of Held *et al.* (2005) is introduced, which is then sequentially improved by suitable model extensions. The final Section 5.4 illustrates simulation from fitted `hhh4` models.

5.1. Model class: `hhh4`

An endemic-epidemic multivariate time-series model for infectious disease counts Y_{it} from units $i = 1, \ldots, I$ during periods $t = 1, \ldots, T$ was proposed by Held *et al.* (2005) and was later extended in a series of papers (Paul *et al.* 2008; Paul and Held 2011; Held and Paul 2012; Meyer and Held 2014a). In its most general formulation, this so-called "`hhh4`" model assumes that, conditional on past observations, Y_{it} has a negative binomial distribution with mean

$$\mu_{it} = e_{it}\,\nu_{it} + \lambda_{it}\,Y_{i,t-1} + \phi_{it}\sum_{j\neq i} w_{ji}\,Y_{j,t-1} \tag{10}$$

and overdispersion parameter $\psi_i > 0$ such that the conditional variance of Y_{it} is $\mu_{it}(1+\psi_i\mu_{it})$. Shared overdispersion parameters, e.g., $\psi_i \equiv \psi$, are supported as well as replacing the negative binomial by a Poisson distribution, which corresponds to the limit $\psi_i \equiv 0$.

Similar to the point process models of Sections 3 and 4, the mean (10) decomposes additively into endemic and epidemic components. The endemic mean is usually modelled proportional to an offset of expected counts e_{it}. In spatial applications of the multivariate `hhh4` model as in this paper, the "unit" i refers to a geographical region and we typically use (the fraction of) the population living in region i as the endemic offset. The observation-driven epidemic component splits up into autoregressive effects, i.e., reproduction of the disease within region i, and neighbourhood effects, i.e., transmission from other regions j. Overall, Equation 10 becomes a rich regression model by allowing for log-linear predictors in all three components:

$$\log(\nu_{it}) = \alpha_i^{(\nu)} + {\boldsymbol{\beta}^{(\nu)}}^\top \boldsymbol{z}_{it}^{(\nu)}\,, \tag{11}$$

$$\log(\lambda_{it}) = \alpha_i^{(\lambda)} + {\boldsymbol{\beta}^{(\lambda)}}^\top \boldsymbol{z}_{it}^{(\lambda)}\,, \tag{12}$$

$$\log(\phi_{it}) = \alpha_i^{(\phi)} + {\boldsymbol{\beta}^{(\phi)}}^\top \boldsymbol{z}_{it}^{(\phi)}\,. \tag{13}$$

The intercepts of these predictors can be assumed identical across units, unit-specific, or random (and possibly correlated). The regression terms often involve sine-cosine effects of time to reflect seasonally varying incidence, but may, e.g., also capture heterogeneous vaccination

[4] `https://survstat.rki.de`

coverage (Herzog *et al.* 2011). Data on infections imported from outside the study region may enter the endemic component (Geilhufe *et al.* 2014), which generally accounts for cases not directly linked to other observed cases, e.g., due to edge effects.

For a single time series of counts Y_t, `hhh4` can be regarded as an extension of `glm.nb` from package **MASS** (Venables and Ripley 2002) to account for autoregression. See the `vignette("hhh4")` for examples of modeling univariate and bivariate count time series using `hhh4`. With multiple regions, spatio-temporal dependence is adopted by the third component in Equation 10 with weights w_{ji} reflecting the flow of infections from region j to region i. These transmission weights may be informed by movement network data (Paul *et al.* 2008; Schrödle *et al.* 2012; Geilhufe *et al.* 2014), but may also be estimated parametrically. A suitable choice to reflect epidemiological coupling between regions (Keeling and Rohani 2008, Chapter 7) is a power-law distance decay $w_{ji} = o_{ji}^{-d}$ defined in terms of the adjacency order o_{ji} in the neighbourhood graph of the regions (Meyer and Held 2014a). Note that we usually normalize the transmission weights such that $\sum_i w_{ji} = 1$, i.e., the $Y_{j,t-1}$ cases are distributed among the regions proportionally to the j'th row vector of the weight matrix (w_{ji}).

Likelihood inference for the above multivariate time-series model has been established by Paul and Held (2011) with extensions for parametric neighbourhood weights by Meyer and Held (2014a). Supplied with the analytical score function and Fisher information, the function `hhh4` by default uses the quasi-Newton algorithm available through the R function `nlminb` to maximize the log-likelihood. Convergence is usually fast even for a large number of parameters. If the model contains random effects, the penalized and marginal log-likelihoods are maximized alternately until convergence. Computation of the marginal Fisher information is accelerated using the **Matrix** package (Bates and Maechler 2015).

5.2. Data structure: `sts`

We briefly introduce the `S4`-class `sts` used for data input in `hhh4` models. See Höhle and Mazick (2010) and Salmon *et al.* (2016) for more detailed descriptions of this class, which is also used for the prospective aberration detection facilities of the **surveillance** package.

The epidemic modeling of multivariate count time series essentially involves three data matrices: a $T \times I$ matrix of the observed counts, a corresponding matrix with potentially time-varying population numbers (or fractions), and an $I \times I$ neighbourhood matrix quantifying the coupling between the I units. In our example, the latter consists of the adjacency orders o_{ji} between the districts. A map of the districts in the form of a `SpatialPolygons` object (defined by the **sp** package) can be used to derive the matrix of adjacency orders automatically using the functions `poly2adjmat` and `nbOrder`, which wrap functionality of package **spdep** (Bivand and Piras 2015):

```
R> weserems_nbOrder <- nbOrder(poly2adjmat(map), maxlag = 10)
```

Given the aforementioned ingredients, the `sts` object `data("measlesWeserEms")` included in **surveillance** has been constructed as follows:

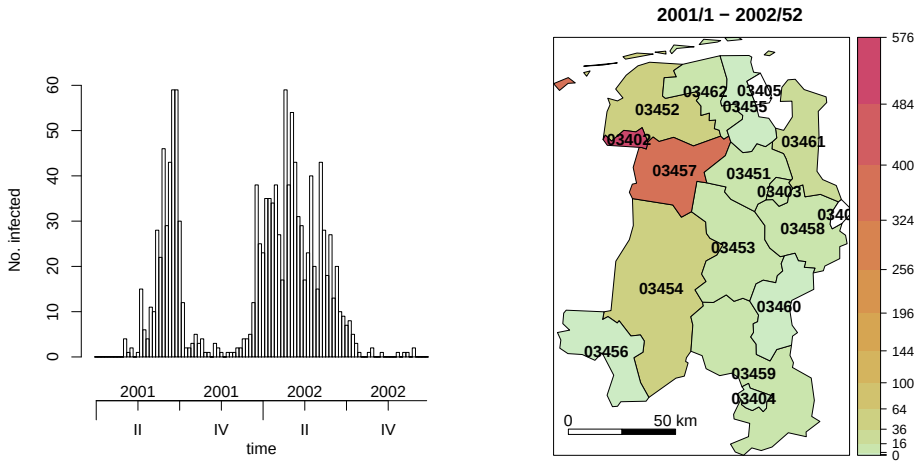
```
R> measlesWeserEms <- sts(observed = counts, start = c(2001, 1), frequency = 52,
+    neighbourhood = weserems_nbOrder, map = map, population = populationFrac)
```

Here, `start` and `frequency` have the same meaning as for classical time-series objects of class `ts`, i.e., (year, sample number) of the first observation and the number of observa-

tions per year. Note that `data("measlesWeserEms")` constitutes a corrected version of `data("measles.weser")` originally used by Held *et al.* (2005).

We can visualize such `sts` data in four ways: individual time series, overall time series, map of accumulated counts by district, or animated maps. For instance, the two plots in Figure 13 have been generated by the following code:

```
R> plot(measlesWeserEms, type = observed ~ time)
R> plot(measlesWeserEms, type = observed ~ unit,
+    population = measlesWeserEms@map$POPULATION / 100000,
+    labels = list(font = 2), colorkey = list(space = "right"),
+    sp.layout = layout.scalebar(measlesWeserEms@map, corner = c(0.05, 0.05),
+      scale = 50, labels = c("0", "50 km"), height = 0.03))
```



(a) Time series of weekly counts.

(b) Disease incidence (per 100 000 inhabitants).

Figure 13: Measles infections in the Weser-Ems region, 2001–2002.

The overall time-series plot in Figure 13a reveals strong seasonality in the data with slightly different patterns in the two years. The spatial plot in Figure 13b is a tweaked `spplot` (package **sp**) with colors from **colorspace** (Ihaka, Murrell, Hornik, Fisher, and Zeileis 2015) using $\sqrt{}$-equidistant cut points handled by package **scales** (Wickham 2015). The default plot `type` is `observed ~ time | unit` and shows the individual time series by district (Figure 14):

```
R> plot(measlesWeserEms, units = which(colSums(observed(measlesWeserEms)) > 0))
```

The plot excludes the districts 03401 (SK Delmenhorst) and 03405 (SK Wilhelmshaven) without any reported cases. Obviously, the districts have been affected by measles to a very heterogeneous extent during these two years.

An animation of the data can be easily produced as well. We recommend to use converters of the **animation** package, e.g., to watch the series of plots in a web browser. The following code will generate weekly disease maps during the year 2001 with the respective total number of cases shown in a legend and – if package **gridExtra** (Auguie 2015) is available – an evolving time-series plot at the bottom:

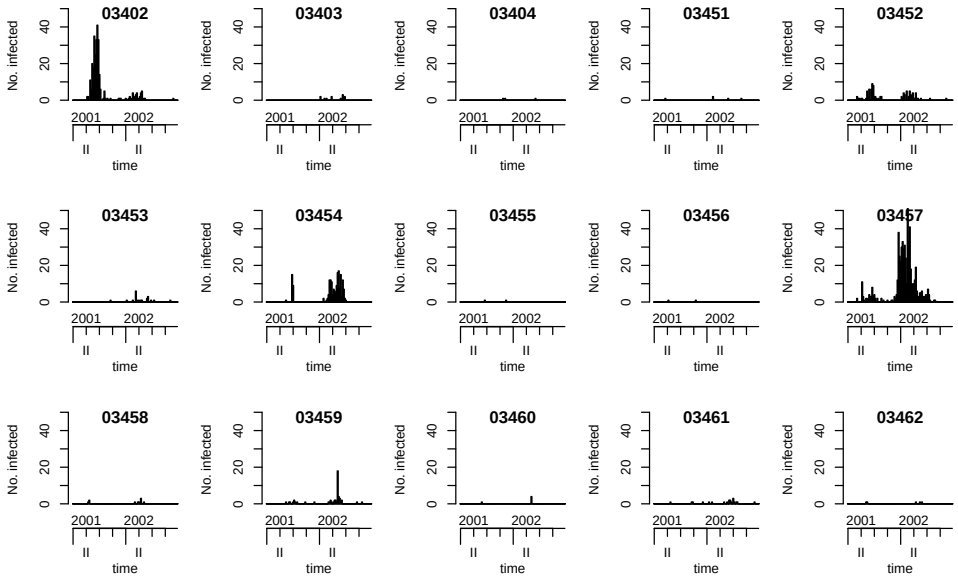


Figure 14: Count time series of the 15 affected districts.

```
R> animation::saveHTML(
+    animate(measlesWeserEms, tps = 1:52, total.args = list()),
+    title = "Evolution of the measles epidemic in the Weser-Ems region, 2001",
+    ani.width = 500, ani.height = 600)
```

5.3. Modeling and inference

For multivariate surveillance time series of counts such as the `measlesWeserEms` data, the function `hhh4` fits models of the form (10) via (penalized) maximum likelihood. We start by modeling the measles counts in the Weser-Ems region by a slightly simplified version of the original negative binomial model by Held *et al.* (2005). Instead of district-specific intercepts $\alpha_i^{(\nu)}$ in the endemic component, we first assume a common intercept $\alpha^{(\nu)}$ in order to not be forced to exclude the two districts without any reported cases of measles. After the estimation and illustration of this basic model, we will discuss the following sequential extensions: covariates (district-specific vaccination coverage), estimated transmission weights, and random effects to eventually account for unobserved heterogeneity of the districts.

Basic model

Our initial model has the following mean structure:

$$\mu_{it} = e_i\,\nu_t + \lambda\,Y_{i,t-1} + \phi\sum_{j\neq i} w_{ji}Y_{j,t-1}\,, \tag{14}$$

$$\log(\nu_t) = \alpha^{(\nu)} + \beta_t t + \gamma\sin(\omega t) + \delta\cos(\omega t)\,. \tag{15}$$

To account for temporal variation of disease incidence, the endemic log-linear predictor ν_t incorporates an overall trend and a sinusoidal wave of frequency $\omega = 2\pi/52$. As a basic district-specific measure of disease incidence, the population fraction e_i is included as a multiplicative offset. The epidemic parameters $\lambda = \exp(\alpha^{(\lambda)})$ and $\phi = \exp(\alpha^{(\phi)})$ are assumed homogeneous across districts and constant over time. Furthermore, we define $w_{ji} = \mathbb{1}(j \sim i) = \mathbb{1}(o_{ji} = 1)$ for the time being, which means that the epidemic can only arrive from directly adjacent districts. This `hhh4` model transforms into the following list of `control` arguments:

```
R> measlesModel_basic <- list(
+    end = list(f = addSeason2formula(~1 + t, period = measlesWeserEms@freq),
+               offset = population(measlesWeserEms)),
+    ar = list(f = ~1),
+    ne = list(f = ~1, weights = neighbourhood(measlesWeserEms) == 1),
+    family = "NegBin1")
```

The formulae of the three predictors $\log \nu_t$, $\log \lambda$ and $\log \phi$ are specified as element `f` of the `end`, `ar`, and `ne` lists, respectively. For the endemic formula we use the convenient function `addSeason2formula` to generate the sine-cosine terms, and we take the multiplicative `offset` of population fractions e_i from the `measlesWeserEms` object. The autoregressive part only consists of the intercept $\alpha^{(\lambda)}$, whereas the neighbourhood component specifies the intercept $\alpha^{(\phi)}$ and also the matrix of transmission `weights` (w_{ji}) to use – here a simple indicator of first-order adjacency. The chosen `family` corresponds to a negative binomial model with a common overdispersion parameter ψ for all districts. Alternatives are `"Poisson"`, `"NegBinM"` (ψ_i), or a factor determining which groups of districts share a common overdispersion parameter. Together with the data, the complete list of control arguments is then fed into the `hhh4` function to estimate the model, a summary of which is printed below.

```
R> measlesFit_basic <- hhh4(stsObj = measlesWeserEms, control = measlesModel_basic)
R> summary(measlesFit_basic, idx2Exp = TRUE, amplitudeShift = TRUE, maxEV = TRUE)

Call:
hhh4(stsObj = measlesWeserEms, control = measlesModel_basic)

Coefficients:
                     Estimate  Std. Error
exp(ar.1)             0.64540   0.07927
exp(ne.1)             0.01581   0.00420
exp(end.1)            1.08025   0.27884
exp(end.t)            1.00119   0.00426
end.A(2 * pi * t/52)  1.16423   0.19212
end.s(2 * pi * t/52) -0.63436   0.13350
overdisp              2.01384   0.28544

Epidemic dominant eigenvalue:  0.72

Log-likelihood:   -971.7
AIC:              1957
BIC:              1996

Number of units:        17
Number of time points:  103
```

The `idx2Exp` argument requests the estimates for λ, ϕ, $\alpha^{(\nu)}$ and $\exp(\beta_t)$ instead of their respective internal log-values. For instance, `exp(end.t)` represents the seasonality-adjusted factor by which the basic endemic incidence increases per week. The `amplitudeShift` argument transforms the internal coefficients γ and δ of the sine-cosine terms to the amplitude A and phase shift φ of the corresponding sinusoidal wave $A\sin(\omega t + \varphi)$ in $\log \nu_t$ (Paul *et al.* 2008). The multiplicative effect of seasonality on ν_t is shown in Figure 15 produced by:

```
R> plot(measlesFit_basic, type = "season", components = "end", main = "")
```

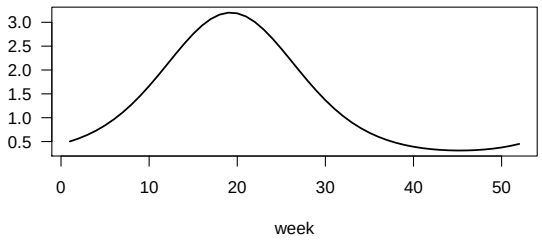


Figure 15: Estimated multiplicative effect of seasonality on the endemic mean.

The `overdisp` parameter and its 95% confidence interval obtained by

```
R> confint(measlesFit_basic, parm = "overdisp")
```

```
         2.5 % 97.5 %
overdisp 1.454  2.573
```

suggest that a negative binomial distribution with overdispersion is more adequate than a Poisson model corresponding to $\psi = 0$. We can underpin this finding by an AIC comparison, taking advantage of the convenient `update` method for `hhh4` fits:

```
R> AIC(measlesFit_basic, update(measlesFit_basic, family = "Poisson"))
```

```
                                          df  AIC
measlesFit_basic                           7 1957
update(measlesFit_basic, family = "Poisson")  6 2479
```

The epidemic potential of the process as determined by the parameters λ and ϕ is best investigated by a combined measure: the dominant eigenvalue (`maxEV`) of the matrix $\mathbf{\Lambda}$ which has the entries $(\Lambda)_{ii} = \lambda$ on the diagonal and $(\Lambda)_{ij} = \phi w_{ji}$ for $j \neq i$ (Paul *et al.* 2008). If the dominant eigenvalue is smaller than unity, it can be interpreted as the epidemic proportion of disease incidence. In the above model, the estimate is 72%. Another way of judging the relative importance of the three model components is to plot the fitted mean components along with the observed counts. Figure 16 shows this for the six districts with more than 20 cases:

```
R> districts2plot <- which(colSums(observed(measlesWeserEms)) > 20)
R> plot(measlesFit_basic, type = "fitted", units = districts2plot, hide0s = TRUE)
```

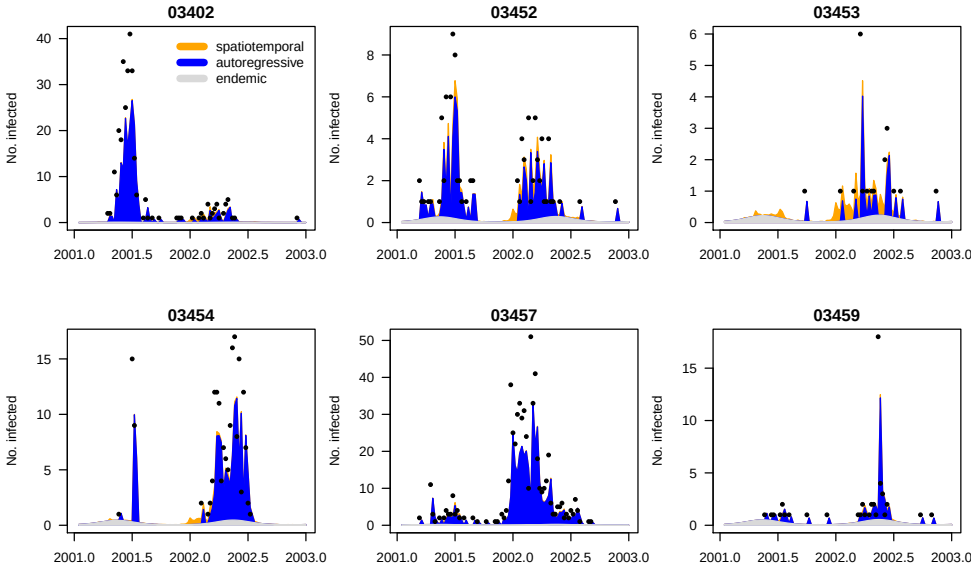


Figure 16: Fitted components in the initial model `measlesFit_basic` for the six districts with more than 20 cases. Dots are only drawn for positive weekly counts.

The largest portion of the fitted mean indeed results from the within-district autoregressive component with very little contribution of cases from adjacent districts and a rather small endemic incidence.

Other plot `type`s and methods for fitted `hhh4` models as listed in Table 8 will be applied in the course of the following model extensions.

Covariates

The `hhh4` model framework allows for covariate effects on the endemic or epidemic contributions to disease incidence. Covariates may vary over both regions and time and thus obey the same $T \times I$ matrix structure as the observed counts. For infectious disease models, the regional vaccination coverage is an important example of such a covariate, since it reflects the (remaining) susceptible population. In a thorough analysis of measles occurrence in the German federal states, Herzog *et al.* (2011) found vaccination coverage to be associated with outbreak size. We follow their approach of using the district-specific proportion $1 - v_i$ of unvaccinated children just starting school as a proxy for the susceptible population. As v_i we use the proportion of children vaccinated with at least one dose among the ones presenting their vaccination card at school entry in district i in the year 2004.[5] This time-constant covariate needs to be transformed to the common matrix structure for incorporation in `hhh4`:

```
R> Sprop <- matrix(1 - measlesWeserEms@map@data$vacc1.2004,
+     nrow = nrow(measlesWeserEms), ncol = ncol(measlesWeserEms), byrow = TRUE)
```

[5]First year with data for all districts – available from the public health department of Lower Saxony (`http://www.nlga.niedersachsen.de/portal/live.php?navigation_id=36791&article_id=135436&_psmand=20`).

Display	Extract	Modify	Other
`print`	`nobs`	`update`	`predict`
`summary`	`coef`		`simulate`
`plot`	`fixef`		`pit`
	`ranef`		`scores`
	`vcov`		`calibrationTest`
	`confint`		`all.equal`
	`coeflist`		*`oneStepAhead`*
	`logLik`		
	`residuals`		
	`terms`		
	`formula`		

Table 8: Generic and *non-generic* functions applicable to `hhh4` objects.

```
R> summary(Sprop[1, ])

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0306  0.0481  0.0581  0.0675  0.0830  0.1400
```

There are several ways to account for the susceptible proportion in our model, among which the simplest is to update the endemic population offset e_i by multiplication with $(1 - v_i)$. Herzog *et al.* (2011) found that the susceptible proportion is best added as a covariate in the autoregressive component in the form

$$\lambda_i \, Y_{i,t-1} = \exp\left(\alpha^{(\lambda)} + \beta_s \log(1 - v_i)\right) Y_{i,t-1} = \exp\left(\alpha^{(\lambda)}\right) (1 - v_i)^{\beta_s} \, Y_{i,t-1}$$

according to the mass action principle (Keeling and Rohani 2008). A higher proportion of susceptibles in district i is expected to boost the generation of new infections, i.e., $\beta_s > 0$. Alternatively, this effect could be assumed as an offset, i.e., $\beta_s \equiv 1$. To choose between endemic and/or autoregressive effects, and multiplicative offset vs. covariate modeling, we perform AIC-based model selection. First, we set up a grid of all combinations of envisaged extensions for the endemic and autoregressive components:

```
R> Soptions <- c("unchanged", "Soffset", "Scovar")
R> SmodelGrid <- expand.grid(end = Soptions, ar = Soptions)
R> row.names(SmodelGrid) <- do.call("paste", c(SmodelGrid, list(sep = "|")))
```

Then we update the initial model `measlesFit_basic` according to each row of `SmodelGrid`:

```
R> measlesFits_vacc <- apply(X = SmodelGrid, MARGIN = 1, FUN = function (options) {
+    updatecomp <- function (comp, option) switch(option,
+      "unchanged" = list(),
+      "Soffset" = list(offset = comp$offset * Sprop),
+      "Scovar" = list(f = update(comp$f, ~. + log(Sprop))))
+    update(measlesFit_basic,
+      end = updatecomp(measlesFit_basic$control$end, options[1]),
+      ar = updatecomp(measlesFit_basic$control$ar, options[2]),
+      data = list(Sprop = Sprop))
+    })
```

The resulting object `measlesFits_vacc` is a list of 9 `hhh4` fits, which are named according to the corresponding `Soptions` used for the endemic and autoregressive component. We construct a call of the function `AIC` taking all list elements as arguments:

```
R> aics_vacc <- do.call(AIC, lapply(names(measlesFits_vacc), as.name),
+    envir = as.environment(measlesFits_vacc))
R> aics_vacc[order(aics_vacc[, "AIC"]), ]

                     df  AIC
`Scovar|unchanged`    8 1917
`Scovar|Scovar`       9 1919
`Soffset|unchanged`   7 1922
`Soffset|Scovar`      8 1924
`Scovar|Soffset`      8 1934
`Soffset|Soffset`     7 1937
unchanged|unchanged   7 1957
`unchanged|Scovar`    8 1959
`unchanged|Soffset`   7 1967
```

Hence, AIC increases if the susceptible proportion is only added to the autoregressive component, but we see a remarkable improvement when adding it to the endemic component. The best model is obtained by leaving the autoregressive component unchanged (λ) and adding the term $\beta_s \log(1 - v_i)$ to the endemic predictor in Equation 15.

```
R> measlesFit_vacc <- measlesFits_vacc[["Scovar|unchanged"]]
R> coef(measlesFit_vacc, se = TRUE)["end.log(Sprop)", ]

  Estimate Std. Error
    1.7181     0.2877
```

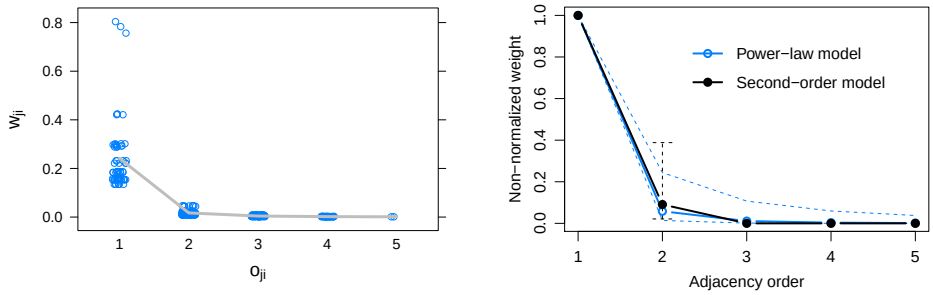
The estimated exponent $\hat{\beta}_s$ is both clearly positive and different from the offset assumption. In other words, if a district's fraction of susceptibles is doubled, the endemic measles incidence is estimated to multiply by $2^{\hat{\beta}_s} = 3.29$ (95% CI: 2.23–4.86).

Spatial interaction

Up to now, the model assumed that the epidemic can only arrive from directly adjacent districts because $w_{ji} = \mathbb{1}(j \sim i)$, and that all districts have the same potential ϕ for importing cases from neighbouring regions. Given the ability of humans to travel further and preferrably to metropolitan areas, both assumptions seem overly simplistic. First, to reflect commuter-driven spread in our model, we scale the district's susceptibility according to its population fraction by multiplying ϕ by $e_i^{\beta_{pop}}$:

```
R> measlesFit_nepop <- update(measlesFit_vacc,
+    ne = list(f = ~log(pop)), data = list(pop = population(measlesWeserEms)))
```

As in a similar analysis of influenza (Meyer and Held 2014a), we find strong evidence for such an agglomeration effect: the estimated exponent is $\hat{\beta}_{pop} = 2.85$ (95% CI: 1.83–3.87) and AIC decreases from 1917 to 1887. Models where attraction to a region scales with population size are called "gravity" models (Xia, Bjørnstad, and Grenfell 2004).



(a) Normalized weights in the power-law model. (b) Non-normalized weights with 95% CIs.

Figure 17: Estimated weights as a function of adjacency order.

To account for long-range transmission of cases, Meyer and Held (2014a) proposed to estimate the weights w_{ji} as a function of the adjacency order o_{ji} between the districts. For instance, a power-law model assumes the form $w_{ji} = o_{ji}^{-d}$, for $j \neq i$ and $w_{jj} = 0$, where the decay parameter d is to be estimated. Normalization to $w_{ji}/\sum_k w_{jk}$ is recommended and applied by default when supplying `W_powerlaw` as weights in the neighbourhood component:

```
R> measlesFit_powerlaw <- update(measlesFit_nepop,
+    ne = list(weights = W_powerlaw(maxlag = 5)))
```

The argument `maxlag` sets an upper bound for spatial interaction in terms of adjacency order. Here we set no limit since `max(neighbourhood(measlesWeserEms))` is 5. The resulting parameter estimate is $\hat{d} = 4.10$ (95% CI: 2.03–6.17), which represents a strong decay of spatial interaction for higher-order neighbours. As an alternative to the parametric power law, unconstrained weights up to `maxlag` can be estimated by using `W_np` instead of `W_powerlaw`. For instance, `W_np(maxlag = 2)` corresponds to a second-order model, i.e., $w_{ji} = 1 \cdot \mathbb{1}(o_{ji} = 1) + e^{\omega_2} \cdot \mathbb{1}(o_{ji} = 2)$, which is also row-normalized by default:

```
R> measlesFit_np2 <- update(measlesFit_nepop,
+    ne = list(weights = W_np(maxlag = 2)))
```

Figure 17b shows both the power law model $o^{-\hat{d}}$ and the second-order model, where $e^{\hat{\omega}_2} = 0.09$ (95% CI: 0.02–0.39). Alternatively, the plot `type = "neweights"` for `hhh4` fits can produce a `stripplot` (Sarkar 2008) of w_{ji} against o_{ji} as shown in Figure 17a for the power-law model:

```
R> library("lattice")
R> plot(measlesFit_powerlaw, type = "neweights", plotter = stripplot,
+    panel = function (...) {panel.stripplot(...); panel.average(...)},
+    jitter.data = TRUE, xlab = expression(o[ji]), ylab = expression(w[ji]))
```

Note that only horizontal jitter is added in this case. Because of normalization, the weight w_{ji} for transmission from district j to district i is determined not only by the districts' neighbourhood o_{ji} but also by the total amount of neighbourhood of district j in the form of $\sum_{k \neq j} o_{jk}^{-d}$, which causes some variation of the weights for a specific order of adjacency.
An AIC comparison of the different models for the transmission weights yields:

```
R> AIC(measlesFit_nepop, measlesFit_powerlaw, measlesFit_np2)

                    df  AIC
measlesFit_nepop     9 1887
measlesFit_powerlaw 10 1882
measlesFit_np2      10 1881
```

AIC improves when accounting for transmission between higher-order neighbours by a power law or a second-order model. In spite of the latter resulting in a slightly better fit, we will use the power-law model as a basis for further model extensions since the stand-alone second-order effect is not always identifiable in more complex models and is scientifically implausible.

Random effects

Paul and Held (2011) introduced random effects for `hhh4` models, which are useful if the districts exhibit heterogeneous incidence levels not explained by observed covariates, and especially if the number of districts is large. For infectious disease surveillance data, a typical example of unobserved heterogeneity is under-reporting (Bernard, Werber, and Höhle 2014). Our measles data even contain two districts without any reported cases, while the district with the smallest population (03402, SK Emden) had the second-largest number of cases reported and the highest overall incidence (see Figures 13b and 14). Hence, allowing for district-specific intercepts in the endemic or epidemic components is expected to improve the model fit. For independent random effects $\alpha_i^{(\nu)} \overset{iid}{\sim} \mathrm{N}(\alpha^{(\nu)}, \sigma_\nu^2)$, $\alpha_i^{(\lambda)} \overset{iid}{\sim} \mathrm{N}(\alpha^{(\lambda)}, \sigma_\lambda^2)$, and $\alpha_i^{(\phi)} \overset{iid}{\sim} \mathrm{N}(\alpha^{(\phi)}, \sigma_\phi^2)$ in all three components, we update the corresponding formulae as follows:

```
R> measlesFit_ri <- update(measlesFit_powerlaw,
+    end = list(f = update(formula(measlesFit_powerlaw)$end, ~. + ri() - 1)),
+    ar  = list(f = update(formula(measlesFit_powerlaw)$ar,  ~. + ri() - 1)),
+    ne  = list(f = update(formula(measlesFit_powerlaw)$ne,  ~. + ri() - 1)))

R> summary(measlesFit_ri, amplitudeShift = TRUE, maxEV = TRUE)

Call:
hhh4(stsObj = object$stsObj, control = control)

Random effects:
            Var   Corr
ar.ri(iid)  1.076
ne.ri(iid)  1.294 0
end.ri(iid) 1.312 0    0

Fixed effects:
                      Estimate  Std. Error
ar.ri(iid)           -1.61389   0.38197
ne.log(pop)           3.42406   1.07722
ne.ri(iid)            6.62429   2.81553
end.t                 0.00578   0.00480
end.A(2 * pi * t/52)  1.20359   0.20149
end.s(2 * pi * t/52) -0.47916   0.14205
end.log(Sprop)        1.79350   0.69159
```

```
end.ri(iid)             4.42260   1.94605
neweights.d             3.60640   0.77602
overdisp                0.97723   0.15132

Epidemic dominant eigenvalue:  0.84

Penalized log-likelihood:  -868.6
Marginal log-likelihood:   -54.2

Number of units:        17
Number of time points:  103
```

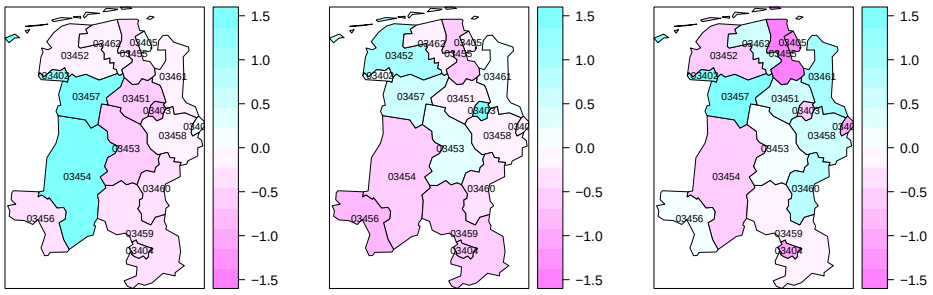
The summary now contains an extra section with the estimated variance components σ^2_λ, σ^2_ϕ, and σ^2_ν of the random effects. We did not assume correlation between the three intercepts, but this is possible by specifying `ri(corr = "all")` in the component formulae. The implementation also supports a conditional autoregressive formulation (Besag, York, and Mollié 1991) for spatially correlated intercepts by using `ri(type = "car")`. The estimated district-specific intercepts can be extracted by the `ranef`-method:

```
R> head(ranef(measlesFit_ri, tomatrix = TRUE), n = 3)

      ar.ri(iid) ne.ri(iid) end.ri(iid)
03401     0.0000   -0.05673     -1.0045
03402     1.2235    0.04312      1.5264
03403    -0.8273    1.55878     -0.6199
```

They can also be visualized in a map by the plot `type = "ri"`:

```
R> for (comp in c("ar", "ne", "end")) {
+    print(plot(measlesFit_ri, type = "ri", component = comp,
+      col.regions = rev(cm.colors(100)), labels = list(cex = 0.6),
+      at = seq(-1.6, 1.6, length.out = 15)))
+  }
```



(a) Autoregressive $\alpha_i^{(\lambda)}$ (b) Spatio-temporal $\alpha_i^{(\phi)}$ (c) Endemic $\alpha_i^{(\nu)}$

Figure 18: Maps of the estimated random intercepts.

For the autoregressive component in Figure 18a, we see a pronounced heterogeneity between the three western districts in blue and the remaining districts. These three districts have been affected by large local outbreaks and are also the ones with the highest overall numbers of cases. In contrast, the city of Oldenburg (03403) is estimated with a relatively low autoregressive factor $\lambda_i = \exp(\alpha^{(\lambda)} + \alpha_i^{(\lambda)}) = 0.0871$, but it seems to import more cases from other districts than explained by its population (Figure 18b). In Figure 18c, the two districts without any reported measles cases (03401 and 03405) appear in dark pink, which means that they exhibit a relatively low endemic incidence after adjusting for the population and susceptible proportion. Such districts could be suspected of a larger amount of under-reporting.

Note that the extra flexiblility of the random effects model comes at a price. First, the estimation runtime increases considerably from 0.1 seconds for the previous power-law model `measlesFit_powerlaw` to 4 seconds with additional random effects. Furthermore, we no longer obtain AIC values in the model summary, since random effects invalidate simple AIC-based model comparisons (Greven and Kneib 2010). Of course we can plot the fitted values and visually compare their quality with the initial fit shown in Figure 16:

```
R> plot(measlesFit_ri, type = "fitted", units = districts2plot, hide0s = TRUE)
```

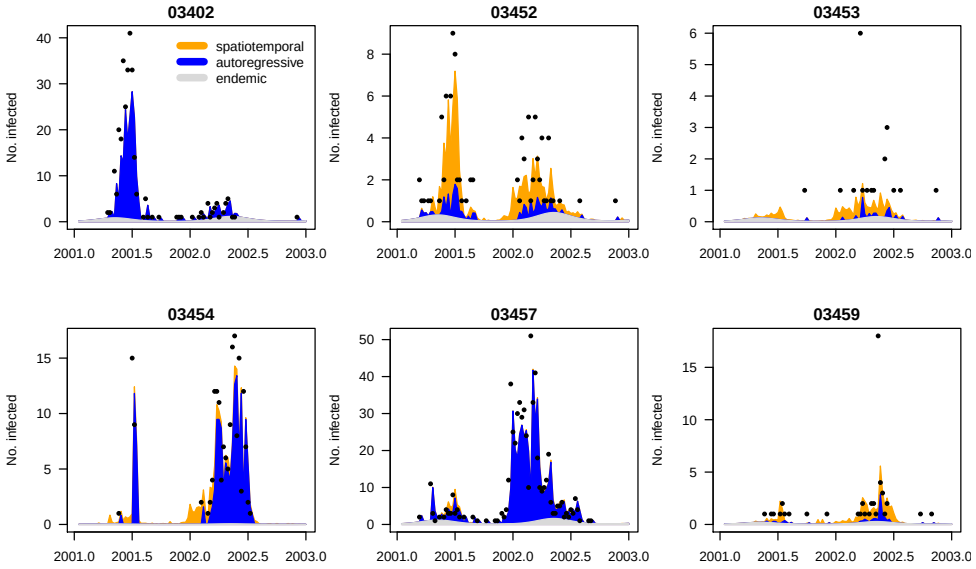


Figure 19: Fitted components in the random effects model `measlesFit_ri` for the six districts with more than 20 cases. Compare to Figure 16.

For some of these districts, a great amount of cases is now explained via transmission from neighbouring regions while others are mainly influenced by the local autoregression. Note that the estimated decomposition of the mean by district can also be seen from the related plot `type = "maps"` (not shown). However, for quantitative comparisons of model performance we have to resort to more sophisticated techniques presented in the next section.

Predictive model assessment

Paul and Held (2011) suggest to evaluate one-step-ahead forecasts from competing models by proper scoring rules for count data (Czado, Gneiting, and Held 2009). These scores measure the discrepancy between the predictive distribution P from a fitted model and the later observed value y. A well-known example is the squared error score ("ses") $(y - \mu_P)^2$, which is usually averaged over a suitable set of forecasts to obtain the mean squared error. More elaborate scoring rules such as the logarithmic score ("logs") or the ranked probability score ("rps") take into account the whole predictive distribution to assess calibration and sharpness simultaneously – see the recent review by Gneiting and Katzfuss (2014). The so-called Dawid-Sebastiani score ("dss") is another option. Lower scores correspond to better predictions.

In the `hhh4` framework, predictive model assessment is made available by the functions `oneStepAhead`, `scores`, `pit`, and `calibrationTest`. We will use the second quarter of 2002 as the test period, and compare the basic model, the power-law model, and the random effects model. First, we use the `"final"` fits on the complete time series to compute the predictions, which then simply correspond to the fitted values during the test period:

```
R> tp <- c(65, 77)
R> models2compare <- paste0("measlesFit_", c("basic", "powerlaw", "ri"))
R> measlesPreds1 <- lapply(mget(models2compare), oneStepAhead,
+    tp = tp, type = "final")
```

Note that in this case, the log-score for a model's prediction in district i in week t equals the associated negative log-likelihood contribution. Comparing the mean scores from different models is thus essentially a goodness-of-fit assessment:

```
R> SCORES <- c("logs", "rps", "dss", "ses")
R> measlesScores1 <- lapply(measlesPreds1, scores, which = SCORES, individual = TRUE)
R> t(sapply(measlesScores1, colMeans, dims = 2))

                     logs    rps    dss   ses
measlesFit_basic    1.089 0.7358 1.2911 5.289
measlesFit_powerlaw 1.101 0.7307 2.2223 5.394
measlesFit_ri       1.007 0.6381 0.9656 4.823
```

All scoring rules claim that the random effects model gives the best fit during the second quarter of 2002. Now we turn to true one-week-ahead predictions of `type = "rolling"`, which means that we always refit the model up to week t to get predictions for week $t + 1$:

```
R> measlesPreds2 <- lapply(mget(models2compare), oneStepAhead,
+    tp = tp, type = "rolling", which.start = "final",
+    cores = 2 * (.Platform$OS.type == "unix"))
R> measlesScores2 <- lapply(measlesPreds2, scores, which = SCORES, individual = TRUE)
R> t(sapply(measlesScores2, colMeans, dims = 2))

                     logs    rps   dss   ses
measlesFit_basic    1.102 0.7478 1.339 5.404
measlesFit_powerlaw 1.136 0.7654 2.929 5.865
measlesFit_ri       1.110 0.7632 2.349 7.080
```

Thus, the most parsimonious initial model `measlesFit_basic` gives the best one-week-ahead predictions in terms of overall mean scores. Statistical significance of the differences in mean scores can be investigated by a `permutationTest` for paired data or a paired t-test:

```
R> set.seed(321)
R> sapply(SCORES, function (score) permutationTest(
+    measlesScores2$measlesFit_ri[, , score],
+    measlesScores2$measlesFit_basic[, , score]))

            logs     rps     dss    ses
diffObs     0.007822 0.01541 1.01   1.677
pVal.permut 0.8669   0.7197  0.5183 0.19
pVal.t      0.8541   0.7165  0.3737 0.1711
```

Hence, there is no clear evidence for a difference between the basic and the random effects model with regard to predictive performance during the test period. Whether predictions of a particular model are well calibrated can be formally investigated by `calibrationTest`s for count data as recently proposed by (Wei and Held 2014). For example:

```
R> calibrationTest(measlesPreds2[["measlesFit_ri"]], which = "rps")

Calibration Test for Count Data (based on RPS)

data:  measlesPreds2[["measlesFit_ri"]]
z = 0.80671, n = 221, p-value = 0.4198
```

Thus, there is no evidence of miscalibrated predictions from the random effects model. Czado *et al.* (2009) describe an alternative informal approach to assess calibration: probability integral transform (PIT) histograms for count data (Figure 20).

```
R> for (m in models2compare)
+    pit(measlesPreds2[[m]], plot = list(ylim = c(0, 1.25), main = m))
```

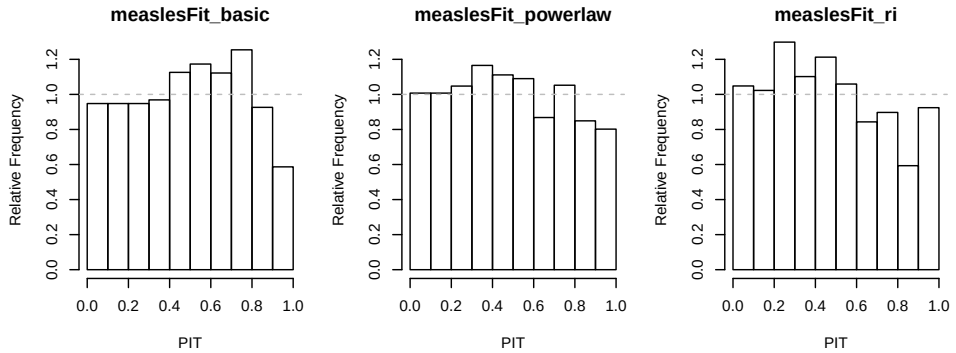


Figure 20: PIT histograms of competing models to check calibration of the one-week-ahead predictions during the second quarter of 2002.

Under the hypothesis of calibration, i.e., $y_{it} \sim P_{it}$ for all predictive distributions P_{it} in the test period, the PIT histogram is uniform. Underdispersed predictions lead to U-shaped histograms, and bias causes skewness. In this aggregate view of the predictions over all districts and weeks of the test period, predictive performance is comparable between the models, and there is no evidence of badly dispersed predictions. However, the right-hand decay in all histograms suggests that all models tend to predict higher counts than observed. This is most likely related to the seasonal shift between the years 2001 and 2002. In 2001, the peak of the epidemic was in the second quarter, while it already occured in the first quarter in 2002 (cp. Figure 13a).

Further modeling options

In the previous sections we extended our model for measles in the Weser-Ems region with respect to spatial variation of the counts and their interaction. Temporal variation was only accounted for in the endemic component, which included a long-term trend and a sinusoidal wave on the log-scale. Held and Paul (2012) suggest to also allow seasonal variation of the epidemic force by adding a superposition of S harmonic waves of fundamental frequency ω, $\sum_{s=1}^{S} \{\gamma_s \sin(s\,\omega t) + \delta_s \cos(s\,\omega t)\}$, to the log-linear predictors of the autoregressive and/or neighbourhood component – just like for $\log \nu_t$ in Equation 15 with $S = 1$. However, given only two years of measles surveillance and the apparent shift of seasonality with regard to the start of the outbreak in 2002 compared to 2001, more complex seasonal models are likely to overfit the data. Concerning the coding in R, sine-cosine terms can be added to the epidemic components without difficulties by again using the convenient function `addSeason2formula`. Updating a previous model for different numbers of harmonics is even simpler, since the `update`-method has a corresponding argument `S`. The plots of `type = "season"` and `type = "maxEV"` for `hhh4` fits can visualize the estimated component seasonality.

All of our models for the measles surveillance data incorporated an epidemic effect of the counts from the local district and its neighbours. Without further notice, we thereby assumed a lag equal to the observation interval of one week. However, the generation time of measles is around 10 days (Anderson and May 1991), which is why some studies, e.g., Finkenstädt, Bjørnstad, and Grenfell (2002) or Herzog *et al.* (2011), aggregate their weekly measles surveillance data into biweekly intervals. Fine and Clarkson (1982) used weekly counts in their analysis and report that biweekly aggregation would have little effect on the results. We can also perform such a sensitivity analysis by running the whole code of the current section based on `aggregate(measlesWeserEms, nfreq = 26)`. Doing so, the parameter estimates of the various models retain their order of magnitude and conclusions remain the same. However, with the number of time points halved, the complex random effects model would not always be identifiable when calculating one-week-ahead predictions during the test period.

We have shown several options to account for the spatio-temporal dynamics of infectious disease spread. However, for directly transmitted human diseases, the social phenomenon of "like seeks like" results in contact patterns between subgroups of a population, which extend the pure distance decay of interaction. Especially for school children, social contacts are known to be highly assortative with respect to age (Mossong *et al.* 2008). A useful epidemic model should therefore be additionally stratified by age group and take the inherent contact structure into account. How this extension can be incorporated in the spatio-temporal endemic-epidemic modeling framework `hhh4` is the focus of current research (Meyer and Held 2015).

5.4. Simulation

Simulation from fitted `hhh4` models is enabled by an associated `simulate`-method. Compared to the point process models of Sections 3 and 4, simulation is less complex since it essentially consists of sequential calls of `rnbinom` (or `rpois`). At each time point t, the mean μ_{it} is determined by plugging in the parameter estimates and the counts $Y_{i,t-1}$ simulated at the previous time point. In addition to a model fit, we thus need to specify an initial vector of counts `y.start`. As an example, we simulate 100 realizations of the evolution of measles during the year 2002 based on the fitted random effects model and the counts of the last week of the year 2001 in the 17 districts:

```
R> (y.start <- observed(measlesWeserEms)[52, ])

03401 03402 03403 03404 03405 03451 03452 03453 03454 03455 03456 03457 03458 03459
    0     0     0     0     0     0     0     0     0     0     0    25     0     0
03460 03461 03462
    0     0     0

R> measlesSim <- simulate(measlesFit_ri,
+    nsim = 100, seed = 1, subset = 53:104, y.start = y.start)
```

The simulated counts are returned as a $52 \times 17 \times 100$ array instead of a list of 100 `sts` objects. We can, e.g., look at the final size distribution of the simulations:

```
R> summary(colSums(measlesSim, dims = 2))

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    223     326     424     550     582    3970
```

A few large outbreaks have been simulated, but the mean size is below the observed number of `sum(observed(measlesWeserEms)[53:104, ])` $= 779$ cases in the year 2002. Using the `plot`-method associated with such `hhh4` simulations, Figure 21 shows the weekly number of observed cases compared to the long-term forecast:

```
R> plot(measlesSim, "time", ylim = c(0, 100))
```

We refer to `help("simulate.hhh4")` for further examples.

6. Conclusion

In the present work we have introduced the R package **surveillance** as a comprehensive statistical framework for the analysis of spatio-temporal surveillance data covering individual-level event data as well as aggregated count data time series. The package offers a multitude of methods for visualization, likelihood inference and simulation of endemic-epidemic models. Additional functionality beyond the illustrations in Sections 3 to 5 can be found via `help(package = "surveillance")`. By the open-source implementation of recently developed statistical methodology in a readily available R package, we support reproducibility of research and hope to serve an increased need in analyzing spatio-temporal epidemic data using statistical models.

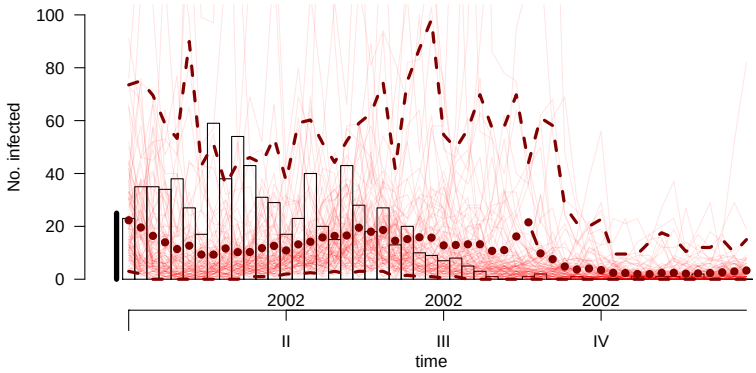


Figure 21: Simulation-based long-term forecast starting from the last week in 2001 (vertical bar on the left), showing the counts aggregated over all districts. The weekly mean of the simulations is represented by dots and the dashed lines correspond to the pointwise 2.5% and 97.5% quantiles. The actually observed counts are shown in the background.

Acknowledgements

The implementation of the `hhh4` model is mainly due to Michaela Paul, to whom we are thankful for all methodological advances and code contributions in the past years. We also acknowledge all other code contributors in the long history of the **surveillance** package (in alphabetical order): Thais Correa, Mathias Hofmann, Christian Lang, Juliane Manitz, Andrea Riebler, Daniel Sabanés Bové, Maëlle Salmon, Dirk Schumacher, Stefan Steiner, Mikko Virtanen, Wei Wei, Valentin Wimmer. Many have also helped us by investigating the package and giving feedback: Doris Altmann, Johannes Dreesman, Johannes Elias, Marc Geilhufe, Jim Hester, Kurt Hornik, Mayeul Kauffmann, Marcos Prates, Brian D. Ripley, Barry Rowlingson, Christopher W. Ryan, Klaus Stark, Yann Le Strat, André Michael Toschke, Wei Wei, George Wood, Achim Zeileis, Bing Zhang. We appreciate the helpful comments from two anonymous reviewers of an earlier version of this manuscript.

Financial support by the Munich Center of Health Sciences (2007–2010) and the Swiss National Science Foundation (2007–2015) is gratefully acknowledged.

Involved R packages and versions

This paper is based on **surveillance** 1.11.0 (Höhle, Meyer, and Paul 2016) in R version 3.2.3 (2015-12-10) using **knitr** (Xie 2015) for dynamic report generation. The implementations of the three presented endemic-epidemic modeling frameworks rely on several other R packages. In the following we list all packages involved as first-order dependencies of **surveillance** with the versions used in this paper: **sp** 1.2-2 (Pebesma and Bivand 2005), **xtable** 1.8-2 (Dahl 2016), **polyCub** 0.5-2 (Meyer 2015), **MASS** 7.3-44 (Venables and Ripley 2002), **Matrix** 1.2-3 (Bates and Maechler 2015), **spatstat** 1.44-1 (Baddeley *et al.* 2015), **lattice** 0.20-33 (Sarkar 2008), **colorspace** 1.2-6 (Ihaka *et al.* 2015), **scales** 0.3.0 (Wickham 2015), **quadprog** 1.5-5 (Turlach 2013), **memoise** 1.0.0 (Wickham *et al.* 2016), **polyclip** 1.3-2 (Johnson 2015), **maptools** 0.8-39

(Bivand and Lewin-Koh 2016) and **spdep** 0.5-92 (Bivand and Piras 2015).

R itself, the **surveillance** package, and all other aforementioned packages are available from the Comprehensive R Archive Network (CRAN) at `https://CRAN.R-project.org/`. The development of **surveillance** is hosted at `http://surveillance.r-forge.r-project.org/`. This manuscript will be turned into a package vignette and kept up-to-date with the software.

References

Adelfio G, Chiodi M (2015). "FLP Estimation of Semi-parametric Models for Space–time Point Processes and Diagnostic Tools." *Spatial Statistics*, **14, Part B**, 119–132. `doi:10.1016/j.spasta.2015.06.004`.

Anderson RM, May RM (1991). *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press.

Auguie B (2015). ***gridExtra**: Miscellaneous Functions for "Grid" Graphics*. R package version 2.0.0, URL `https://CRAN.R-project.org/package=gridExtra`.

Baddeley A, Rubak E, Turner R (2015). *Spatial Point Patterns: Methodology and Applications with R*. Chapman and Hall/CRC Press, London. In press, URL `http://www.crcpress.com/Spatial-Point-Patterns-Methodology-and-Applications-with-R/Baddeley-Rubak-Turner/9781482210200/`.

Balderama E, Schoenberg FP, Murray E, Rundel PW (2012). "Application of Branching Models in the Study of Invasive Species." *Journal of the American Statistical Association*, **107**(498), 467–476. `doi:10.1080/01621459.2011.641402`.

Bates D, Maechler M (2015). ***Matrix**: Sparse and Dense Matrix Classes and Methods*. R package version 1.2-3, URL `https://CRAN.R-project.org/package=Matrix`.

Bernard H, Werber D, Höhle M (2014). "Estimating the Under-Reporting of Norovirus Illness in Germany Utilizing Enhanced Awareness of Diarrhoea During a Large Outbreak of Shiga Toxin-Producing E. Coli O104:H4 in 2011 – A Time Series Analysis." *BMC Infectious Diseases*, **14**(1), 116. `doi:10.1186/1471-2334-14-116`.

Besag J, York J, Mollié A (1991). "Bayesian Image-Restoration, with Two Applications in Spatial Statistics." *The Annals of the Institute of Statistical Mathematics*, **43**(1), 1–20.

Bivand R, Keitt T, Rowlingson B (2015). ***rgdal**: Bindings for the Geospatial Data Abstraction Library*. R package version 1.1-3, URL `https://CRAN.R-project.org/package=rgdal`.

Bivand R, Lewin-Koh N (2016). ***maptools**: Tools for Reading and Handling Spatial Objects*. R package version 0.8-39, URL `https://CRAN.R-project.org/package=maptools`.

Bivand R, Piras G (2015). "Comparing Implementations of Estimation Methods for Spatial Econometrics." *Journal of Statistical Software*, **63**(18), 1–36. URL `http://www.jstatsoft.org/v63/i18/`.

Bivand RS, Pebesma E, Gómez-Rubio V (2013). *Applied Spatial Data Analysis with R*, volume 10 of *Use R!* 2nd edition. Springer-Verlag, New York. ISBN 1-4614-7617-8. URL http://www.asdar-book.org.

Brown PE (2015). "Model-Based Geostatistics the Easy Way." *Journal of Statistical Software*, **63**(12), 1–24. URL http://www.jstatsoft.org/v63/i12.

Cori A, Ferguson NM, Fraser C, Cauchemez S (2013). "A New Framework and Software to Estimate Time-Varying Reproduction Numbers During Epidemics." *American Journal of Epidemiology*, **178**(9), 1505–1512. doi:10.1093/aje/kwt133.

Czado C, Gneiting T, Held L (2009). "Predictive Model Assessment for Count Data." *Biometrics*, **65**(4), 1254–1261. doi:10.1111/j.1541-0420.2009.01191.x.

Dahl DB (2016). ***xtable**: Export Tables to LaTeX or HTML*. R package version 1.8-2, URL https://CRAN.R-project.org/package=xtable.

Daley DJ, Gani J (1999). *Epidemic Modelling: An Introduction*, volume 15 of *Cambridge Studies in Mathematical Biology*. Cambridge University Press. ISBN 0-521-64079-2. doi:10.1017/CBO9780511608834.

Daley DJ, Vere-Jones D (2003). *An Introduction to the Theory of Point Processes*, volume I: Elementary Theory and Methods of *Probability and its Applications*. 2nd edition. Springer-Verlag, New York. ISBN 0-387-95541-0.

Diggle PJ (2006). "Spatio-Temporal Point Processes, Partial Likelihood, Foot and Mouth Disease." *Statistical Methods in Medical Research*, **15**(4), 325–336. doi:10.1191/0962280206sm454oa.

Douglas DH, Peucker TK (1973). "Algorithms for the Reduction of the Number of Points Required to Represent a Digitized Line or its Caricature." *Cartographica: The International Journal for Geographic Information and Geovisualization*, **10**(2), 112–122. doi:10.3138/FM57-6770-U75U-7727.

Fahrmeir L, Kneib T, Lang S, Marx B (2013). *Regression: Models, Methods and Applications*. Springer-Verlag. ISBN 3-642-34332-5. doi:10.1007/978-3-642-34333-9.

Fine PEM, Clarkson JA (1982). "Measles in England and Wales—I: An Analysis of Factors Underlying Seasonal Patterns." *International Journal of Epidemiology*, **11**(1), 5–14. doi:10.1093/ije/11.1.5.

Finkenstädt BF, Bjørnstad ON, Grenfell BT (2002). "A Stochastic Model for Extinction and Recurrence of Epidemics: Estimation and Inference for Measles Outbreaks." *Biostatistics*, **3**(4), 493–510. doi:10.1093/biostatistics/3.4.493.

Finkenstädt BF, Grenfell BT (2000). "Time Series Modelling of Childhood Diseases: A Dynamical Systems Approach." *Journal of the Royal Statistical Society C*, **49**(2), 187–205. doi:10.1111/1467-9876.00187.

Geilhufe M, Held L, Skrøvseth SO, Simonsen GS, Godtliebsen F (2014). "Power Law Approximations of Movement Network Data for Modeling Infectious Disease Spread." *Biometrical Journal*, **56**(3), 363–382. doi:10.1002/bimj.201200262.

Gneiting T, Katzfuss M (2014). "Probabilistic Forecasting." *Annual Review of Statistics and Its Application*, **1**(1), 125–151. doi:10.1146/annurev-statistics-062713-085831.

Greven S, Kneib T (2010). "On the Behaviour of Marginal and Conditional AIC in Linear Mixed Models." *Biometrika*, **97**(4), 773–789. doi:10.1093/biomet/asq042.

Groendyke C, Welch D, Hunter DR (2012). "A Network-Based Analysis of the 1861 Hagelloch Measles Data." *Biometrics*, **68**(3), 755–765. doi:10.1111/j.1541-0420.2012.01748.x.

Harrower M, Bloch M (2006). "Mapshaper.org: A Map Generalization Web Service." *IEEE Computer Graphics and Applications*, **26**(4), 22–27. doi:10.1109/MCG.2006.85.

Held L, Hofmann M, Höhle M, Schmid V (2006). "A Two-Component Model for Counts of Infectious Diseases." *Biostatistics*, **7**(3), 422–437. doi:10.1093/biostatistics/kxj016.

Held L, Höhle M, Hofmann M (2005). "A Statistical Framework for the Analysis of Multivariate Infectious Disease Surveillance Counts." *Statistical Modelling*, **5**(3), 187–199. doi:10.1191/1471082X05st098oa.

Held L, Paul M (2012). "Modeling Seasonality in Space-Time Infectious Disease Surveillance Data." *Biometrical Journal*, **54**(6), 824–843. doi:10.1002/bimj.201200037.

Herzog SA, Paul M, Held L (2011). "Heterogeneity in Vaccination Coverage Explains the Size and Occurrence of Measles Epidemics in German Surveillance Data." *Epidemiology and Infection*, **139**(04), 505–515. doi:10.1017/S0950268810001664.

Hughes AW, King ML (2003). "Model Selection Using AIC in the Presence of One-Sided Information." *Journal of Statistical Planning and Inference*, **115**(2), 397–411. doi:10.1016/S0378-3758(02)00159-3.

Höhle M (2007). "**surveillance**: An R Package for the Monitoring of Infectious Diseases." *Computational Statistics*, **22**(4), 571–582. doi:10.1007/s00180-007-0074-8.

Höhle M (2009). "Additive-Multiplicative Regression Models for Spatio-Temporal Epidemics." *Biometrical Journal*, **51**(6), 961–978. doi:10.1002/bimj.200900050.

Höhle M (2016). "Infectious Disease Modelling." In AB Lawson, S Banerjee, RP Haining, MD Ugarte (eds.), *Handbook of Spatial Epidemiology*, Chapman & Hall/CRC Handbooks of Modern Statistical Methods. Chapman and Hall/CRC. ISBN 1-4822-5301-1. Forthcoming, http://www.math.su.se/~hoehle/pubs/Hoehle_SpaMethInfEpiModelling2015.pdf.

Höhle M, Mazick A (2010). "Aberration Detection in R Illustrated by Danish Mortality Monitoring." In TA Kass-Hout, X Zhang (eds.), *Biosurveillance: Methods and Case Studies*, pp. 215–238. Chapman and Hall/CRC.

Höhle M, Meyer S, Paul M (2016). ***surveillance**: Temporal and Spatio-Temporal Modeling and Monitoring of Epidemic Phenomena*. R package version 1.11.0, URL http://surveillance.r-forge.r-project.org/.

Höhle M, Paul M, Held L (2009). "Statistical Approaches to the Monitoring and Surveillance of Infectious Diseases for Veterinary Public Health." *Preventive Veterinary Medicine*, **91**(1), 2–10. doi:10.1016/j.prevetmed.2009.05.017.

Ihaka R, Murrell P, Hornik K, Fisher JC, Zeileis A (2015). ***colorspace**: Color Space Manipulation*. R package version 1.2-6, URL http://CRAN.R-project.org/package=colorspace.

Johnson A (2015). ***polyclip**: Polygon Clipping*. R package version 1.3-2, ported to R by Adrian Baddeley and Brian Ripley, URL https://CRAN.R-project.org/package=polyclip.

Johnson SD (2010). "A Brief History of the Analysis of Crime Concentration." *European Journal of Applied Mathematics*, **21**(Special Double Issue 4-5), 349–370. doi:10.1017/S0956792510000082.

Jombart T, Cori A, Didelot X, Cauchemez S, Fraser C, Ferguson N (2014). "Bayesian Reconstruction of Disease Outbreaks by Combining Epidemiologic and Genomic Data." *PLOS Computational Biology*, **10**(1), e1003457. doi:10.1371/journal.pcbi.1003457.

Keeling MJ, Rohani P (2008). *Modeling Infectious Diseases in Humans and Animals*. Princeton University Press. ISBN 0-691-11617-2. URL http://www.modelinginfectiousdiseases.org/.

Kermack WO, McKendrick AG (1927). "A Contribution to the Mathematical Theory of Epidemics." *Proceedings of the Royal Society of London A*, **115**(772), 700–721. doi:10.1098/rspa.1927.0118.

Lawson AB, Leimich P (2000). "Approaches to the Space-Time Modelling of Infectious Disease Behaviour." *IMA Journal of Mathematics Applied in Medicine and Biology*, **17**(1), 1–13.

Liboschik T, Fokianos K, Fried R (2015). "**tscount**: An R Package for Analysis of Count Time Series Following Generalized Linear Models." *SFB 823 Discussion Paper 6/2015*, TU Dortmund. URL http://CRAN.R-project.org/package=tscount.

Malesios C, Demiris N, Kalogeropoulos K, Ntzoufras I (2014). "Bayesian Spatio-Temporal Epidemic Models with Applications to Sheep Pox." http://arxiv.org/abs/1403.1783.

Martinussen T, Scheike TH (2002). "A Flexible Additive Multiplicative Hazard Model." *Biometrika*, **89**(2), 283–298. doi:10.1093/biomet/89.2.283.

Merl D, Johnson LR, Gramacy RB, Mangel M (2010). "**amei**: An R Package for the Adaptive Management of Epidemiological Interventions." *Journal of Statistical Software*, **36**(6), 1–32. URL http://www.jstatsoft.org/v36/i06.

Meyer S (2015). ***polyCub**: Cubature over Polygonal Domains*. R package version 0.5-2, URL http://CRAN.R-project.org/package=polyCub.

Meyer S, Elias J, Höhle M (2012). "A Space-Time Conditional Intensity Model for Invasive Meningococcal Disease Occurrence." *Biometrics*, **68**(2), 607–616. doi:10.1111/j.1541-0420.2011.01684.x. http://arxiv.org/abs/1508.05740.

Meyer S, Held L (2014a). "Power-Law Models for Infectious Disease Spread." *The Annals of Applied Statistics*, **8**(3), 1612–1639. doi:10.1214/14-AOAS743.

Meyer S, Held L (2014b). "Supplement B of 'Power-Law Models for Infectious Disease Spread'." doi:10.1214/14-AOAS743SUPPB. http://www.biostat.uzh.ch/research/manuscripts/powerlaw.html.

Meyer S, Held L (2015). "Incorporating Social Contact Data in Spatio-Temporal Models for Infectious Disease Spread." http://arxiv.org/abs/1512.01065.

Meyer S, Warnke I, Rössler W, Held L (2015). "Model-Based Testing for Space-Time Interaction Using Point Processes: An Application to Psychiatric Hospital Admissions in an Urban Area." *Spatial and Spatio-temporal Epidemiology*. In revision, http://arxiv.org/abs/1512.09052.

Mohler GO, Short MB, Brantingham PJ, Schoenberg FP, Tita GE (2011). "Self-exciting Point Process Modeling of Crime." *Journal of the American Statistical Association*, **106**(493), 100–108. doi:10.1198/jasa.2011.ap09546.

Mossong J, Hens N, Jit M, Beutels P, Auranen K, Mikolajczyk R, Massari M, Salmaso S, Tomba GS, Wallinga J, Heijne J, Sadkowska-Todys M, Rosinska M, Edmunds WJ (2008). "Social Contacts and Mixing Patterns Relevant to the Spread of Infectious Diseases." *PLoS Medicine*, **5**(3), e74. doi:10.1371/journal.pmed.0050074.

Neal PJ, Roberts GO (2004). "Statistical Inference and Model Selection for the 1861 Hagelloch Measles Epidemic." *Biostatistics*, **5**(2), 249–261. doi:10.1093/biostatistics/5.2.249.

Obadia T, Haneef R, Boelle PY (2012). "The **R0** Package: A Toolbox to Estimate Reproduction Numbers for Epidemic Outbreaks." *BMC Medical Informatics and Decision Making*, **12**(147). doi:10.1186/1472-6947-12-147.

Ogata Y (1988). "Statistical Models for Earthquake Occurrences and Residual Analysis for Point Processes." *Journal of the American Statistical Association*, **83**(401), 9–27. URL http://www.jstor.org/stable/2288914.

Ogata Y (1999). "Seismicity Analysis Through Point-process Modeling: A Review." *Pure and Applied Geophysics*, **155**(2), 471–507. doi:10.1007/s000240050275.

Paul M, Held L (2011). "Predictive Assessment of a Non-Linear Random Effects Model for Multivariate Time Series of Infectious Disease Counts." *Statistics in Medicine*, **30**(10), 1118–1136. doi:10.1002/sim.4177.

Paul M, Held L, Toschke A (2008). "Multivariate Modelling of Infectious Disease Surveillance Data." *Statistics in Medicine*, **27**(29), 6250–6267. doi:10.1002/sim.3440.

Pebesma E (2012). "**spacetime**: Spatio-Temporal Data in R." *Journal of Statistical Software*, **51**(7), 1–30. URL http://www.jstatsoft.org/v51/i07/.

Pebesma E (2016). "CRAN Task View: Handling and Analyzing Spatio-Temporal Data." http://CRAN.R-project.org/web/views/SpatioTemporal.html. Version 2016-01-11.

Pebesma EJ, Bivand RS (2005). "Classes and Methods for Spatial Data in R." *R News*, **5**(2), 9–13. URL http://CRAN.R-project.org/doc/Rnews/.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rowlingson B, Diggle P (2015). ***splancs**: Spatial and Space-Time Point Pattern Analysis*. R package version 2.01-38, URL https://CRAN.R-project.org/package=splancs.

Ryan JA, Ulrich JM (2014). ***xts**: eXtensible Time Series*. R package version 0.9-7, URL `https://CRAN.R-project.org/package=xts`.

Salmon M, Schumacher D, Höhle M (2016). "Monitoring Count Time Series in R: Aberration Detection in Public Health Surveillance." *Journal of Statistical Software*. In press, `http://arxiv.org/abs/1411.1292`.

Sarkar D (2008). *Lattice: Multivariate Data Visualization with R*. Springer-Verlag, New York. ISBN 978-0-387-75968-5, URL `http://lmdvr.r-forge.r-project.org`.

Scheike TH, Martinussen T (2006). *Dynamic Regression Models for Survival Data*. Springer-Verlag.

Schrödle B, Held L, Rue H (2012). "Assessing the Impact of a Movement Network on the Spatiotemporal Spread of Infectious Diseases." *Biometrics*, **68**(3), 736–744. `doi:10.1111/j.1541-0420.2011.01717.x`.

Silvapulle MJ, Sen PK (2005). *Constrained Statistical Inference: Order, Inequality, and Shape Constraints*. Wiley Series in Probability and Statistics. John Wiley & Sons. ISBN 0-471-20827-2.

Sommariva A, Vianello M (2007). "Product Gauss Cubature over Polygons based on Green's Integration Formula." *Bit Numerical Mathematics*, **47**(2), 441–453. `doi:10.1007/s10543-007-0131-2`.

Stadler T, Bonhoeffer S (2013). "Uncovering Epidemiological Dynamics in Heterogeneous Host Populations Using Phylogenetic Methods." *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, **368**(1614), 20120198. `doi:10.1098/rstb.2012.0198`.

Therneau TM (2015). *A Package for Survival Analysis in S*. Version 2.38, URL `http://CRAN.R-project.org/package=survival`.

Turlach BA (2013). ***quadprog**: Functions to solve Quadratic Programming Problems*. R package version 1.5-5, ported to R by Andreas Weingessel, URL `https://CRAN.R-project.org/package=quadprog`.

Utsu T, Ogata Y, Matsu'ura RS (1995). "The Centenary of the Omori Formula for a Decay Law of Aftershock Activity." *Journal of Physics of the Earth*, **43**(1), 1–33. `doi:10.4294/jpe1952.43.1`.

Venables WN, Ripley BD (2002). *Modern Applied Statistics with S*. Fourth edition. Springer-Verlag, New York. ISBN 0-387-95457-0, URL `http://www.stats.ox.ac.uk/pub/MASS4`.

Vrbik I, Deardon R, Feng Z, Gardner A, Braun J (2012). "Using Individual-level Models for Infectious Disease Spread to Model Spatio-temporal Combustion Dynamics." *Bayesian Analysis*, **7**(3), 615–638. `doi:10.1214/12-BA721`.

Waller LA, Gotway CA (2004). *Applied Spatial Statistics for Public Health Data*. Wiley Series in Probability and Statistics. John Wiley & Sons. ISBN 0-471-66267-4. `doi:10.1002/0471662682`.

Wei W, Held L (2014). "Calibration Tests for Count Data." *TEST*, **23**(4), 787–805. doi:10.1007/s11749-014-0380-8.

Wickham H (2015). ***scales**: Scale Functions for Visualization*. R package version 0.3.0, URL https://CRAN.R-project.org/package=scales.

Wickham H, Hester J, Müller K (2016). ***memoise**: Memoisation of Functions*. R package version 1.0.0, URL https://CRAN.R-project.org/package=memoise.

Xia Y, Bjørnstad ON, Grenfell BT (2004). "Measles Metapopulation Dynamics: A Gravity Model for Epidemiological Coupling and Dynamics." *The American Naturalist*, **164**(2), 267–281. URL http://www.jstor.org/stable/10.1086/422341.

Xie Y (2013). "**animation**: An R Package for Creating Animations and Demonstrating Statistical Methods." *Journal of Statistical Software*, **53**(1), 1–27. URL http://www.jstatsoft.org/v53/i01/.

Xie Y (2015). *Dynamic Documents with R and **knitr***. The R Series, 2nd edition. Chapman and Hall/CRC, Boca Raton, Florida. ISBN 1-4987-1696-2. URL http://yihui.name/knitr/.

Affiliation:

Sebastian Meyer
Epidemiology, Biostatistics and Prevention Institute
University of Zurich
Hirschengraben 84
CH-8001 Zurich, Switzerland
E-mail: sebastian.meyer@uzh.ch
URL: http://www.ebpi.uzh.ch/en/aboutus/departments/biostatistics.html

Leonhard Held
Epidemiology, Biostatistics and Prevention Institute
University of Zurich
E-mail: leonhard.held@uzh.ch

Michael Höhle
Department of Mathematics
Stockholm University
E-mail: hoehle@math.su.se
URL: http://www.math.su.se/~hoehle

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
MMMMMM YYYY, Volume VV, Issue II *Submitted:* yyyy-mm-dd
doi:10.18637/jss.v000.i00 *Accepted:* yyyy-mm-dd

PAPER III

Incorporating social contact data in spatio-temporal models for infectious disease spread

Sebastian Meyer, Leonhard Held

Revised for *Biostatistics*, 2016.

Biostatistics (2016), **0**, 0, *pp.* 1–15
doi:10.1093/biostatistics/paper

Incorporating social contact data in spatio-temporal models for infectious disease spread

SEBASTIAN MEYER*, LEONHARD HELD

Epidemiology, Biostatistics and Prevention Institute, University of Zurich, Hirschengraben 84, CH-8001 Zürich, Switzerland

sebastian.meyer@uzh.ch

Summary

Routine public health surveillance of notifiable infectious diseases gives rise to weekly counts of reported cases – possibly stratified by region and/or age group. A well-established approach to the statistical analysis of such surveillance data are endemic-epidemic time-series models. The temporal dependence inherent to communicable diseases is thereby taken into account by an observation-driven formulation conditioning on past counts. Additional spatial dynamics of area-level incidence are largely driven by human travel and can be captured by power-law weights based on the order of adjacency. However, social contacts are highly assortative also with respect to age. For example, characteristic pathways of directly transmitted pathogens are linked to child-care facilities, schools and nursing homes. We therefore investigate how an age-structured social contact matrix can be incorporated into a spatio-temporal endemic-epidemic model for infectious disease counts. To illustrate the approach, we analyze the spread of norovirus gastroenteritis over 6 age groups within the 12 districts of Berlin, 2011–2015, using contact data from the POLYMOD study. The proposed age-structured model outperforms alternative scenarios with homogeneous or no mixing between age groups. An extended contact model suggests a power transformation of the survey-based contact matrix towards more within-group transmission.

Key words: Age-structured contact matrix; Areal count time series; Endemic-epidemic modelling; Infectious disease epidemiology; Norovirus gastroenteritis; Norwalk virus; Spatio-temporal surveillance data.

1. Introduction

The social phenomenon of "like seeks like" produces characteristic contact patterns between certain subgroups of a population. The quantification of such social mixing behaviour to model the spread of infectious diseases in human populations has been put forward by the work of Wallinga *and others* (1999). A recent review of studies measuring epidemiologically relevant social mixing behaviour is provided by Read *and others* (2012). One of the largest surveys to date was conducted as part of the EU-funded POLYMOD project and recorded conversational contacts of 7 290 individuals in eight European countries (Mossong *and others*, 2008). Contact

*To whom correspondence should be addressed.

patterns were found to be similar across the different countries and highly assortative with respect to age, especially for school children and young adults.

The basic idea behind the combination of such social contact data with epidemic models has been termed the "social contact hypothesis" (Wallinga *and others*, 2006): The age-specific numbers of potentially infectious contacts are proportional to age-specific numbers of social contacts. For instance, for pathogens transmitted via respiratory droplets, face-to-face conversation and/or physical contact are frequently used as proxy measures for exposure. Goeyvaerts *and others* (2010) use data on such contacts from the POLYMOD study to estimate the so-called "who acquires infection from whom" (WAIFW) matrix from seroprevalence data. Many other studies have now made use of the POLYMOD contact data as well (Rohani *and others*, 2010; Birrell *and others*, 2011; Baguelin *and others*, 2013; Goeyvaerts *and others*, 2015), but none of them takes the spatial characteristics of disease spread into account. The distance of social contacts from the home location of each participant has only recently been investigated by Read *and others* (2014). Their finding that "most were within a kilometre of the participant's home, while some occurred further than 500 km away" reflects the power-law distance decay of social interaction as determined by human travel behaviour (Carrothers, 1956; Cliff and Ord, 1975; Brockmann *and others*, 2006). Meyer and Held (2014) found such a power law to translate to the spatial spread of infectious diseases.

The purpose of this paper is to combine both the social and spatial determinants of infectious disease spread in a time-series model for public health surveillance data. Such data are routinely available as weekly counts of reported cases of a notifiable disease aggregated across administrative districts and possibly further stratified by age group or gender. Social contact matrices reflect the amount of mixing between these strata. Since a useful statistical model should take this structure into account, we extend our previously established spatio-temporal model for area-level count time series (Meyer and Held, 2014) by an additional stratification variable. In this paper, focus is on age-structured models, but the presented methodology equivalently applies to other strata or combinations thereof, such as age-gender groups. We investigate if incorporating a (possibly adjusted) contact matrix captures disease spread better than more simple assumptions of homogeneous or no mixing between the different groups of the population. The approach also allows us to estimate how much disease incidence in each group can be linked to previous cases in their own and in other groups – while adjusting for the spatial pattern of disease spread.

The structure of this paper is as follows. Section 2 introduces our case study on weekly counts of norovirus gastroenteritis in Berlin, 2011–2015, stratified by city district and age group, as well as the social contact data from the POLYMOD study that we make use of. Section 3 first outlines the spatio-temporal modelling framework and then describes how to incorporate an additional stratification variable featuring a contact matrix. Section 4 shows the results of applying the proposed time-series model to the Berlin norovirus data. We conclude the paper with a discussion in Section 5. Supplementary material contains additional figures, an animation of the data, as well as the data and software to reproduce the presented analysis.

2. Case study: Norovirus gastroenteritis in Berlin, 2011–2015

Most of the aforementioned studies relate contact patterns to the spread of influenza, whereas we here investigate the occurrence of norovirus-associated acute gastroenteritis. Both diseases are highly infectious, have a similar temporal pattern, and similar mortality in elderly persons (van Asten *and others*, 2012). However, in contrast to influenza, vaccines against noroviruses have yet to be developed (Pringle *and others*, 2015). From a statistical modelling perspective, absence of

vaccination simplifies the analysis of infectious disease occurrence since vaccination coverage – potentially varying across age groups, regions and over time – needs not to be taken into account. In the following, we give an overview of the general epidemiology of norovirus gastroenteritis, and describe the surveillance data from Berlin and the social contact data that we use.

2.1 *Epidemiology of norovirus gastroenteritis*

The review of Pringle *and others* (2015) summarizes the current knowledge about the epidemiology of and immunity to noroviruses. Norovirus-associated acute gastroenteritis is characterized by "sudden onset of vomiting, diarrhea and abdominal cramps lasting 2–3 days", but the symptom profile varies by age group. O'Dea *and others* (2014) estimate an average symptomatic period of 3.35 days from several outbreaks in hospitals and long-term care facilities, where vulnerable individuals live closely together and norovirus outbreaks most commonly occur. Another group of the population frequently affected by norovirus gastroenteritis are young children in daycare centres. Norovirus incidence exhibits strong seasonality with a peak during winter, where outbreaks in childcare facilities were observed to precede those in private households, hospitals and nursing homes (Bernard *and others*, 2014).

Noroviruses are highly contagious since only few viral particles are needed for an infection, and they are thermally stable and particularly persistent in the environment (Marshall and Bruggink, 2011). Thus, noroviruses are not only transmitted directly from person to person, but also indirectly via contaminated surfaces or food. Furthermore, fecal norovirus shedding decreases over time but its duration varies by age and is in the range of several weeks. These characteristics suggest a rather large variation in the serial interval, i.e. the time between onset of symptoms in a primary and a secondary case, which ranges from within a day to more than one week with a median of about 3 days estimated for an outbreak across several daycare facilities in Stockholm (Götz *and others*, 2001; Heijne *and others*, 2009). Estimation of the serial interval is complicated by asymptomatic norovirus infections, the proportion of which was reported to be 32% in a volunteer study (Graham *and others*, 1994). These persons can shed virus, albeit less than symptomatic shedders (Atmar *and others*, 2008; Sukhrie *and others*, 2012), and they are invisible to public health surveillance systems.

2.2 *Incidence data*

In Germany, the national public health institute (the Robert Koch Institute, RKI) provides access to incidence data of notifiable diseases through the *SurvStat@RKI 2.0* online service (`https://survstat.rki.de`) allowing for customized queries to the German notification system database. Since the last revision of the case definition for norovirus gastroenteritis in 2011, only laboratory-confirmed cases are reported to the RKI. Previously, data from epidemiologically linked cases of acute gastroenteritis were collected as well, but this procedure has been cancelled because of its unreliability and high burden on local public health departments (Robert Koch-Institut, 2010). To obtain data from a consistent reporting regime, we thus restrict our attention to the time since 2011. Note that the number of *reported* cases to be modelled excludes all asymptomatic cases as well as all those symptomatic cases, who have not found their way to laboratory testing (Gibbons *and others*, 2014). It is known that under-reporting of norovirus illness is most pronounced in the 20 to 29 year-old persons and substantially lower in persons aged < 10 years and 70 years and over (Bernard *and others*, 2014). A sensitivity analysis will indicate how under-reporting affects the interpretation of our model results.

As to the geographic region of interest, we chose the largest city of Germany: Berlin. With respect to RKI's reporting system, Berlin is unique in that it is divided further into 12 administrative districts for each of which disease notification data is separately available. This enables the analysis of disease spread on smaller spatial scales. Furthermore, a large underlying population is required for our time-series model of the number of infections (ignoring the depletion of susceptibles) to be a reasonable approximation of the epidemic process (Farrington *and others*, 2003).

We have downloaded weekly numbers of reported cases of norovirus gastroenteritis in Berlin according to the current reference definition from *SurvStat@RKI* (as of the annual report 2015). The extracted time series cover four norovirus seasons from 2011-W27 to 2015-W26 and are stratified by the 12 city districts and 6 age groups: 0-4, 5-14, 15-24, 25-44, 45-64 and 65+ years of age. The age groups were condensed from original 5-year intervals to reflect heterogeneous epidemiological mixing of pre-school versus school children and intergenerational mixing. Similarly stratified population numbers were obtained from the Statistical Information System Berlin-Brandenburg *StatIS-BBB* (`https://www.statistik-berlin-brandenburg.de/statis`) at the reference date 31 December 2011, where Berlin had 3 501 872 inhabitants in total with age group fractions of 4.6%, 7.8%, 10.4%, 30.4%, 27.9% and 18.9%, respectively. To determine and display spatial dependencies in the areal count time series, the geographical shapes of Berlin's 97 local centres (subsequently aggregated to the 12 city districts) were obtained from the Statistical Office of Berlin-Brandenburg (`http://daten.berlin.de/datensaetze`) as of July 2012.

Figure 1 shows the weekly numbers of reported norovirus gastroenteritis cases aggregated over all city districts and stratified by age group. The mean yearly incidence is 349 (0-4), 23 (5-14), 31 (15-24), 36 (25-44), 52 (45-64) and 266 (65+) cases per 100 000 inhabitants, respectively, so pre-school children and the retired population show a 5 to 15-fold increased incidence compared to the other age groups. The yearly seasonal pattern with overall counts ranging from 7 to 214 cases per week is approximately constant during the four years (supplementary Figure S1). The typical bump during the Christmas break may be related to reporting deficiencies and school closure (Hens *and others*, 2009). The strong peak in the time series of the 5 to 14 year old children is due to a large food-borne outbreak from 20 September to 5 October 2012 (weeks 38 to 40), which affected Berlin and several other states in East Germany. The outbreak has been traced back to contaminated frozen strawberries, which were delivered almost exclusively by one large catering company serving schools and childcare facilities (Bernard *and others*, 2014). A comparison of seasonality between the age groups confirms the findings of Bernard *and others* (2014) from analysing norovirus gastroenteritis occurrence over all of Germany from 2001 to 2009: cases in younger individuals are notified earlier than adult cases. The age-structured spatio-temporal modelling approach presented in Section 3 will help to address the question raised by Bernard *and others* (2014), "whether this reflects a pattern of disease transmission from young to old in the community" – taking the spatial aspect of disease spread into account.

How disease incidence varies across the 12 city districts of Berlin is shown in Figure 2. The south-western district Steglitz-Zehlendorf tends to be affected more and the central districts tend to be affected less than the remaining districts. This pattern is roughly consistent across age groups. An exception are the two younger age groups, which exhibit a relatively high incidence in Marzahn-Hellersdorf. District-specific seasonal shifts are not apparent (supplementary Figure S2).

Animated, age-stratified maps of the weekly norovirus disease counts encompass the full information from all three data dimensions. Such an animation can be found in the supplementary material and may provide additional insight into the dynamics of disease spread. However, endemic-epidemic modelling of the data as in Section 3 provides a more structured view and implicitly takes population heterogeneity into account.

2.3 *Contact data*

To inform our model about the degree of mixing and therefore transmission potential between the different age groups, we use contact data from the German subset of the POLYMOD study (Mossong *and others*, 2008). We count both physical and non-physical (conversational) contacts, but will also report results for a contact matrix based on physical contacts only. The age-structured social contact matrix $\boldsymbol{C} = (c_{g'g})$ contains the mean numbers of contact persons in age group g during one day reported by a participant in age group g'. Instead of using sample means, we estimate the contact matrix by the maximum-likelihood approach of Wallinga *and others* (2006). This assumes a negative binomial distribution for each entry $c_{g'g}$ and takes the reciprocal nature of contacts into account by imposing $c_{g'g}n_{g'} = c_{gg'}n_g$, where n_g is Berlin's population in age group g. We estimate a detailed contact matrix with 5-year intervals, which we subsequently aggregate to the above 6 characteristic age groups (Figure 3). Direct estimation of the aggregated contact matrix leads to similar numbers.

The strong diagonal pattern in the social contact matrix reflects that people tend to mix with people of the same age, which is especially true for school children. The other prominent pattern is produced by the contacts between parents and children. The matrix for physical contacts shows similar patterns (supplementary Figure S3). Aggregation of the contact matrix is done by summing over the contact groups (columns) to be joined and calculating the weighted average across the corresponding participant groups (rows), with weights equal to the group sizes. The aggregated contact matrix is rather asymmetric because of the different sizes of the involved age groups, but reciprocity at the population level still holds. For the models described in the next section, only the row-wise distributions will be relevant, i.e. the contact pattern of an infectious participant across the different age groups. For instance, the contacts of pre-school children mainly belong to the age group 25 to 44 years, but the contact persons of that group mainly belong to itself.

3. An age-structured spatio-temporal model for infectious disease counts

We start this section by reviewing a previously proposed endemic-epidemic modelling framework for aggregated spatio-temporal surveillance data (Meyer and Held, 2014, Section 3). This model is subsequently extended by an additional data dimension to account for the highly assortative mixing of people as quantified by social contact studies.

3.1 *Spatio-temporal formulation*

An endemic-epidemic multivariate time-series model for areal counts Y_{rt} from regions $r = 1, \dots, R$ during periods $t = 1, \dots, T$ was originally proposed by Held *and others* (2005). Conditionally on past observations, Y_{rt} is assumed to follow a negative binomial distribution with mean μ_{rt} and region-specific overdispersion parameters ψ_r such that the conditional variance of Y_{rt} is $\mu_{rt}(1 + \psi_r\mu_{rt})$. The lower bound $\psi_r = 0$ yields the Poisson distribution as a special case, and a common simplifying assumption is that $\psi_r = \psi$ is shared across regions. In its most general formulation, the mean of this spatio-temporal model is additively decomposed into *endemic* and observation-driven *epidemic* components as

$$\mu_{rt} = e_{rt}\,\nu_{rt} + \lambda_{rt}\,Y_{r,t-1} + \phi_{rt}\sum_{r'\neq r}\lfloor w_{r'r}\rfloor\,Y_{r',t-1} \tag{3.1}$$

with log-linear predictors

$$\begin{aligned}\log(\nu_{rt}) &= \alpha_r^{(\nu)} + {\boldsymbol{\beta}^{(\nu)}}^\top \boldsymbol{z}_{rt}^{(\nu)}\,,\\ \log(\lambda_{rt}) &= \alpha_r^{(\lambda)} + {\boldsymbol{\beta}^{(\lambda)}}^\top \boldsymbol{z}_{rt}^{(\lambda)}\,,\\ \log(\phi_{rt}) &= \alpha_r^{(\phi)} + {\boldsymbol{\beta}^{(\phi)}}^\top \boldsymbol{z}_{rt}^{(\phi)}\,,\end{aligned} \tag{3.2}$$

and normalized transmission weights $\lfloor w_{r'r} \rfloor := w_{r'r} / \sum_j w_{r'j}$, $w_{rr} = 0$. In (3.2), the intercepts $\alpha_r^{(\cdot)}$ of the three log-linear predictors can be assumed identical across regions, region-specific, or random (Paul and Held, 2011). The regression terms often involve sine-cosine effects of time to reflect seasonally varying incidence (Held and Paul, 2012), but $\boldsymbol{z}_{rt}^{(\cdot)}$ may also capture other explanatory variables, such as vaccination coverage (Herzog *and others*, 2011). The endemic mean is modelled proportional to an offset of expected counts e_{rt}, for which we typically use (the fraction of) the population living in region r (assumed constant over time). The endemic component partially accounts for cases not directly linked to observed cases from the previous time period, e.g. due to commuting or travelling outside the study region (edge effects).

Disregarding the temporal dimension, the model for the mean endemic incidence is very much related to classical disease mapping approaches for non-contagious spatial processes (Wakefield, 2007). However, since surveillance data of infectious diseases exhibit autoregressive behaviour and occasional outbreaks, a so-called epidemic component driven by the observed counts $(Y_{1,t-1}, \ldots, Y_{R,t-1})$ in the previous period is superposed on the endemic component. The epidemic component in (3.1) splits up into autoregressive effects, i.e. reproduction of the disease within region r, and neighbourhood effects, i.e. transmission from other regions r'. It has proven useful to account for population size also in $\log(\phi_{rt}) = \alpha^{(\phi)} + \tau \log(e_{rt})$ (Meyer and Held, 2014). The parameter τ thereby determines how "attraction" to a region scales with population size (Xia *and others*, 2004). Furthermore, the transmission weights $w_{r'r}$ reflect the flow of infections from region r' to region r. These weights may be based on additional movement network data (Paul *and others*, 2008; Schrödle *and others*, 2012; Geilhufe *and others*, 2014), but may also be estimated from the data at hand. A suitable parametric model to reflect epidemiological coupling between regions is a power-law distance decay $w_{r'r} = o_{r'r}^{-\rho}$ defined in terms of the adjacency order $o_{r'r}$ in the neighbourhood graph of the regions (Meyer and Held, 2014). With such a power law and the population dependence of ϕ_{rt} mentioned above, the neighbour-driven component of (3.1) becomes similar to the gravity model studied by Xia *and others* (2004). The relation between the two models is described in detail in Höhle (2016).

Estimating separate dynamics for the reproduction of the disease within a region on the one hand, and transmission from other regions on the other hand, goes back to the original model formulation of Held *and others* (2005), where only first-order neighbours have been incorporated: $w_{r'r} = \mathbb{1}(o_{r'r} = 1)$. The parametric distance weights offer an appealing alternative to reflect predominant local autoregression in a simpler model with a single epidemic component:

$$\mu_{rt} = e_{rt}\,\nu_{rt} + \phi_{rt} \sum_{r'} \lfloor w_{r'r} \rfloor\, Y_{r',t-1}\,, \tag{3.3}$$

where the choice $w_{r'r} = (o_{r'r} + 1)^{-\rho}$ gives unit weight to local transmission ($r' = r$) and then decays as a power law in terms of adjacency order. Apart from using fewer parameters, this two-component formulation has the advantage of being extensible more naturally to an additional stratification variable, here age group.

3.2 *Extension for stratified areal count time series*

The above model formulation takes into account that people do not have equally distributed contacts across all regions. The social contact structure similarly implies non-homogeneous mixing among different subgroups of the population. Extending the above spatio-temporal model to fit areal time series of counts $Y_{g,r,t}$ stratified by group $g = 1, \ldots, G$ in addition to region, enables us to relax the simple assumption of homogeneous mixing within each region. Our focus in this paper is on age-structured models such that g indexes different age groups. More complex strata such as the interaction of age group and gender are equally possible and can be subsumed in the single group index g. Note, however, that further stratification leads to lower counts per group and region, which might cause identifiability problems in fitting the corresponding models.

We assume that a contact matrix $\boldsymbol{C} = (c_{g'g})$ is given, where each entry $c_{g'g} \geqslant 0$ quantifies the average number of contacts of an individual of group g' with individuals of group g. The spatio-temporal model (3.3) then extends to a three-dimensional version as follows:

$$\mu_{grt} = e_{grt}\,\nu_{grt} + \phi_{grt} \sum_{g',r'} \lfloor c_{g'g}\, w_{r'r} \rfloor\, Y_{g',r',t-1}\,. \tag{3.4}$$

The potentially time-varying population offset e_{grt} is now by group $\times$ region and the endemic and epidemic predictors may gain group-specific effects. How the counts from the previous period affect the current mean in group g and region r is now determined by a product of contact and spatial weights. The product ensures that cases from group g' in region r' are ignored both if there are no contacts to group g or the geographical relation to region r suggests absence of transmission. The weights are row-normalized as in the unstratified model, now over all combinations of group and region: $\sum_{g,r} c_{g'g} w_{r'r} = 1$. Note that this normalization removes any differences in group-specific overall contact rates (the row sums of $\boldsymbol{C}$). Our model therefore does not distinguish between proportionate mixing, where the rows of the contact matrix only differ by a proportionality factor, and a matrix with identical rows. Instead of group-specific *infectiousness*, the endemic-epidemic model estimates group-specific *susceptibility*: The regression term ϕ_{grt} calibrates the weighted sum of past cases transmitted to group g in region r. For example, if $\phi_{grt} = \phi_g^{(G)} \phi_{rt}^{(RT)}$, the group-specific effects $\phi_g^{(G)}$ will adjust the *columns* of the contact matrix.

There are two special cases of the contact structure involved in the epidemic component. First, a contact matrix with identical rows implies that the mixing pattern of the $Y_{g',r',t-1}$ infectious cases does not depend on the group g' they belong to. An example of such homogeneous mixing is a contact matrix where each row contains the G population fractions ($c_{g'g} = e_g$). If ϕ_{grt} contains group-specific effects as mentioned above, a simple matrix of ones ($c_{g'g} = 1$) will induce the same contact structure. The other special case is a diagonal contact matrix $\boldsymbol{C} = \boldsymbol{I}$, which reflects complete absence of mixing. This is equivalent to modelling the G areal time series separately using the spatio-temporal regression framework (3.3). However, also in this case of no between-group mixing, the joint model formulation has the advantage of allowing for parsimonious decompositions of ν_{grt} and ϕ_{grt} into group and region effects. Borrowing strength across groups is especially useful in applications with low counts. In what follows we propose a way to incorporate the given contact matrix $\boldsymbol{C}$ in an adaptive way to infer how much interaction between the groups the surveillance data actually support.

3.3 *Parameterising the contact matrix*

Informing epidemic models by external contact data from sociological studies is a sensible approach. Especially for models of low-prevalence diseases, there might be no feasible alternative to

taking the contact matrix as granted. However, the contact patterns from such surveys might not fully match the characteristics of disease spread. For example, social networks are known to change during illness (van Kerckhove *and others*, 2013) and brief contacts are frequently not reported (Smieszek *and others*, 2014). We therefore suggest a parsimonious single-parameter approach to adaptively estimate the transmission weights as a function of the given contact matrix $\boldsymbol{C}$.

Our proposal for a parametrisation of the contact matrix is borrowed from Küchenhoff *and others* (2006), who deal with misclassification in regression models. A misclassification matrix is progressively transformed to establish an association between the amount of misclassification in a covariate and the corresponding parameter estimate. The involved transformation uses the eigendecomposition of the matrix $\boldsymbol{C}$ to raise it to the power of $\kappa \geqslant 0$,

$$\boldsymbol{C}^{\kappa} := \boldsymbol{E}\boldsymbol{\Lambda}^{\kappa}\boldsymbol{E}^{-1}\,, \tag{3.5}$$

where $\boldsymbol{\Lambda}$ is the diagonal matrix of eigenvalues and $\boldsymbol{E}$ is the corresponding matrix of eigenvectors. Translated to our setting, the parameter κ measures the amount of transmission between the different groups of the population. Specifically, $\kappa = 0$ corresponds to complete absence of between-group transmission ($\boldsymbol{C} = \boldsymbol{I}$), whereas $\kappa = 1$ leaves the social contact matrix unchanged. If $\boldsymbol{C}$ is row-normalized, the behaviour of $\boldsymbol{C}^{\kappa}$ for $\kappa > 1$ can be derived from regarding $\boldsymbol{C}$ as the transition matrix of a homogeneous Markov chain (Grimmett and Stirzaker, 2001, Chapter 6). In particular, for $\kappa \to \infty$, all rows of $\boldsymbol{C}^{\kappa}$ converge to the stationary distribution of the process (if it exists). Thus, with increasing κ the transmission pattern becomes independent of the group the infected individual belongs to. Because of this useful interpretation, we assume a row-normalized contact matrix in the remainder of this paper, but keep the original notation $\boldsymbol{C}$.

Conditions for the existence of $\boldsymbol{C}^{\kappa}$ are briefly discussed in Küchenhoff *and others* (2006, Appendix A). The basic requirement that $\boldsymbol{C}$ can be factorized by an eigendecomposition will hold in most practical cases. However, we also need to make sure that $\boldsymbol{C}^{\kappa}$ has non-negative entries for $\kappa < 1$. In our application, only two entries in $\boldsymbol{C}^{\kappa}$ become negative (close to 0) for small κ. We therefore follow a pragmatic approach and truncate negative entries at 0. See Figure 4b for an illustration of how diagonal and off-diagonal entries, respectively, are affected by raising $\boldsymbol{C}$ to the power of κ. Overall, the parametrisation (3.5) enables an assessment of how far a given contact matrix translates to transmission patterns between the subgroups of the population. Figure 4a exemplifies $\boldsymbol{C}^{\kappa}$ for the row-normalized version of the contact matrix from Figure 3.

3.4 *Inference*

Likelihood inference for the multivariate count time-series model (3.1) has been worked out by Paul and Held (2011), with extensions for parametric neighbourhood weights by Meyer and Held (2014). The log-likelihood is maximized numerically using the quasi-Newton algorithm available through the R function `nlminb` (R Core Team, 2016). Supplied with analytical formulae for the score function and Fisher information, convergence is fast also for a large number of parameters. The associated modelling framework is implemented in the R package `surveillance` (Meyer *and others*, 2016, Section 5) as function `hhh4`.

The extended age-structured model (3.4) is built on top of the existing inference framework, also implementing group-specific transmission weights and group-specific overdispersion. The power parameter κ of (3.5) is conveniently estimated via a profile likelihood approach (see e.g. Held and Sabanés Bové, 2014, Section 5.3), which avoids the cumbersome implementation of additional derivatives with respect to all model parameters. We numerically maximize the log-likelihood of a model with fixed contact matrix $\boldsymbol{C}^{\kappa}$ as a function of κ. The profile confidence interval for κ incorporates the uncertainty of all other parameter estimates (but not vice versa).

4. Results

We present results of applying an age-structured spatio-temporal model of the form (3.4) to the norovirus surveillance data described in Section 2. To account for the heterogeneous numbers of cases per age group, we allow for group-specific overdispersion parameters ψ_g. For the mean, the following endemic-epidemic structure is assumed:

$$\begin{aligned}\mu_{grt} = e_{gr} \exp\left\{\alpha_g^{(G)} + \alpha_r^{(R)} + \beta x_t + \gamma_g \sin(\omega t) + \delta_g \cos(\omega t)\right\} \\ + \phi_g^{(G)} \phi_r^{(R)} e_{gr}^{\tau} \sum_{g',r'} \lfloor (\boldsymbol{C}^{\kappa})_{g'g} \, (o_{r'r}+1)^{-\rho} \rfloor \, Y_{g',r',t-1} \,,\end{aligned} \tag{4.6}$$

for $g = 1, \ldots, 6$ age groups, $r = 1, \ldots, 12$ city districts of Berlin, and $t = 1, \ldots, 208$ weeks with $t = 1$ corresponding to 2011-W27. Note that identifiability of the involved intercepts is in fact ensured by fixing $\alpha_1^{(G)} = \alpha_1^{(R)} = 0$, $\phi_1^{(G)} = \phi_1^{(R)} = 1$, and including overall intercepts in both components. The fraction of the population in age group g and district r, e_{gr}, enters both the endemic component as a multiplicative offset and the epidemic component as a scaling factor. The endemic predictor allows for age- and district-specific incidence levels, fewer cases during the Christmas break ($x_t = 1$ in calendar weeks 52 and 1, otherwise $x_t = 0$), as well as age-specific yearly seasonality ($\omega = 2\pi/52$). Transmission between age groups is modelled using the row-normalized version of the estimated contact matrix from the POLYMOD study, raised to the power of κ via the eigendecomposition (3.5). Transmission between districts is quantified by a power law with respect to adjacency order. The effect of the weighted sum of past counts on the mean μ_{grt} depends on both age group and district via the parameters $\phi_g^{(G)}$ and $\phi_r^{(R)}$ (estimated on the log-scale), respectively. This mainly reflects that certain age groups or districts may be more susceptible to infection than others. Furthermore, these scaling parameters are influenced by the degree of under-reporting, since missing case reports tend to lower the epidemic component.

Table 1 summarizes competing models with respect to the assumed contact structure between age groups. It turns out that a superposed epidemic component greatly improves upon a purely endemic model, and that incorporating the contact matrix from the POLYMOD study outperforms naive models with homogeneous or no mixing between age groups. Akaike's Information Criterion (AIC) is minimal for the model with a power-adjusted contact matrix $\boldsymbol{C}^{\kappa}$ (penultimate row), where the exponent is estimated to be $\hat{\kappa} = 0.47$ (95% CI: 0.34 to 0.66). This means that the epidemic part subsumes more information from cases in the own age group than suggested by the original contact matrix from the POLYMOD study (cf. Figure 4a). The difference in AIC associated with this adjustment, however, is minor compared to the improvement achieved by replacing the naive assumptions on $\boldsymbol{C}$ by the POLYMOD contact matrix in the first place. Results are very similar for the *physical* contact matrix, but the fit is slightly worse.

The spatial spread of the disease across city districts is estimated to have a strong distance decay with $\hat{\rho} = 2.27$ (95% CI: 1.98 to 2.61), i.e. the weights attributed to the adjacency orders 0 to 4 are 1.00, 0.21, 0.08, 0.04 and 0.03. Because human movement patterns are known to depend on age (Read *and others*, 2014; Kucharski *and others*, 2015), we have tried to estimate age-dependent decay parameters replacing ρ by $\rho_{g'}$ in (4.6). However, the parameter for the low-incidence group of 5 to 14 year old children was not identifiable. If we thus assumed a shared decay parameter for both the 5 to 14 and 15 to 24 year old persons, the model becomes identifiable but the resulting estimate still has a large uncertainty. AIC increases for the age-dependent power-law model because the other age groups show very similar decay parameters (supplementary Figure S4). We have assessed the basic power-law assumption by unconstrained estimates of the order-specific weights (Meyer and Held, 2014), which are close to the power law (Figure S4).

In accordance with the idea of a spatio-temporal gravity model (Xia *and others*, 2004), we find that the amount of weekly disease counts linked to cases from the previous week scales with the population size of the "importing" district and age group. Similar to a previous application on influenza (Meyer and Held, 2014), the corresponding estimate $\hat{\tau} = 0.86$ (95% CI: 0.53 to 1.19) is slightly below unity and provides strong evidence for such an association. A model with identical population offsets in both the endemic and epidemic component, i.e. $\tau = 1$, as assumed by Xia *and others* (2004) would provide a similar fit.

To understand how much the weekly disease incidence in the individual age groups can be linked to cases from the previous week, Figure 5 shows the endemic-epidemic decomposition of the estimated mean aggregated across districts (the supplementary Figures S5 and S6 show the district-level fit). When aggregated over all age groups (supplementary Figure S7), the epidemic component explains about half of the weekly disease incidence. Accordingly, when reformulating the model as a multivariate branching process with immigration (Held and Paul, 2012), the largest eigenvalue of the matrix holding the estimated coefficients of $Y_{g',r',t-1}$ is 0.71, which can be interpreted as the overall epidemic proportion of disease incidence. However, this value mostly reflects the situation for the 65+ age group where the within-group spread is dominating. In contrast, for the groups of 5 to 14 and 15 to 24 year old persons, almost no dependence on past counts of the same or the other age groups can be identified. Interestingly, the groups of 25 to 44 and 45 to 64 year old persons seem to inherit a relevant proportion of cases from other age groups with smaller reproduction in their own group. The youngest age group, though, depends on its own cases rather than on cases in other age groups, which is probably related to this age group having the earliest onset of the yearly wave of norovirus infections. The estimated age-dependent sine-cosine effects on the endemic component, which capture these shifts, are shown in supplementary Figure S8 and confirm the findings of Bernard *and others* (2014). The group-specific modal endemic incidence is in calendar weeks 48 (0-4), 45 (5-14), 52 (15-24), 51 (25-44), 52 (45-64) and 3 (65+), respectively. The largest amplitude is estimated for the youngest and oldest age groups.

The estimated group-specific overdispersion parameters are 0.24 (0-4), 1.98 (5-14), 0.30 (15-24), 0.03 (25-44), 0.15 (45-64) and 0.40 (65+) in the AIC-optimal model with adjusted contact matrix. The level of overdispersion varies considerably between age groups and is largest in the group of 5 to 14 year old children. This may be partly due to the relatively strong peak this group experiences during the food-borne outbreak in 2012, for which the model does not explicitly account in the mean structure. The estimated overdispersion parameters are similar between the different epidemic models of Table 1, but slightly larger in the endemic-only model.

5. Discussion

In this paper, we have shown how a social contact matrix can be incorporated in an adaptive way in a regression-oriented endemic-epidemic time-series model for stratified, area-level infectious disease counts. This three-dimensional approach provides more detailed insight into the dynamics of disease spread than unstratified areal models, which inherently assume homogeneous mixing within each spatial unit. Furthermore, a joint model can borrow strength across districts and groups to identify parameters which could not be estimated in separate group-specific areal models, say.

Our application to age-stratified norovirus incidence in Berlin revealed superiority of the contact model compared to naive assumptions of either homogeneous or no mixing between age groups. The model further improved when adjusting the contact matrix from the POLYMOD

study towards more transmission within the own age group. This could be related to biases in contact reporting (Smieszek *and others*, 2014) with more unreported (short) contacts along the diagonal. Overall, the two age groups covering parents were affected the most by preceeding infections in other age groups. This corresponds to the inverse finding of Worby *and others* (2015) that school children played the leading role during influenza A epidemics in that their infections tend to preceed the overall epidemic peak. The estimated strong power-law decay in the spatial dimension means that new infections are predominantly explained by past cases from the same district, with a considerably smaller dependence on districts further apart.

The spatial power law was originally motivated by human travel behaviour on larger scales. An alternative formulation for our city-level data would be a single-step kernel which only discriminates within-district spread from homogeneous transmission to other districts, i.e. $w_{r'r} = 1 \cdot \mathbb{1}(o_{r'r} = 0) + \rho \cdot \mathbb{1}(o_{r'r} > 0)$. Such a kernel provides a comparable fit in our application, but we have preferred the power-law model as a more sensible formulation, in particular for spatial extrapolation. Another option to model spatial dependence comes to mind when interpreting the proposed model as a multidimensional "patch" or metapopulation model with stratification by region as well as age group. In this sense, one could try to replace the parametric power law by a social contact matrix, which is stratified by spatial distance in addition to age group. Separate movement data for school children and adults could then be used to quantify the strength of epidemiological coupling between regions (Kucharski *and others*, 2015). However, integration of movement network data does not necessarily improve predictions (Geilhufe *and others*, 2014).

Social contacts are assortative not only with respect to age and distance. As recently shown by Barclay *and others* (2014), school children even tend to mix according to their vaccination status. A model adaption to more complex contact structures is straightforward provided correspondingly stratified surveillance data is available. A potentially more severe simplification of our model is the assumption of a time-constant contact matrix. Although, weekday vs. weekend differences in contact behaviour are not relevant for weekly time-series models, there are other possibly relevant seasonal effects on larger time scales. Most importantly, the contact structure of school children changes considerably between regular and school holiday periods as was shown in the POLYMOD data (Hens *and others*, 2009). Our model could be further tuned both by using a time-varying contact matrix or by estimating seasonality also in the epidemic component (Held and Paul, 2012), which the `hhh4` implementation already supports.

To check how robust our results are with respect to under-reporting, we re-estimated the models with age-specific multiplication factors applied to the reported numbers of cases. Roughly following Bernard *and others* (2014, Table 1), we used factors of 1.5 (0-4), 2.5 (5-14), 3.0 (15-24), 3.0 (25-44), 2.5 (45-64) and 2.0 (65+), respectively. Except for the naturally increasing overdispersion, the parameter estimates are close to the original fit and the proportions of cases imported from other age groups are similar (supplementary Figure S9). For small strata with a low number of cases, a drawback of this simple deterministic approach is that zero reported counts remain zero regardless of the amount of under-reporting. More sophisticated adjustments for unnotified symptomatic infections are currently investigated within a Bayesian modelling framework. In principle, asymptomatic infections could be similarly accounted for as missing cases, but they seem to play a minor role in disease transmission (Sukhrie *and others*, 2012) and constitute a different issue in public health planning as they do not require treatment. However, one-week-ahead forecasts or long-term simulations of the number of (symptomatic) infections are of particular relevance. Whether the improved model fit with social contact data also leads to better predictions will be described elsewhere.

Supplementary material

The reader is referred to the Supplementary Materials for animated maps of norovirus infections by age group in Berlin, 2011–2015, and additional figures as referenced in Sections 2, 4 and 5. We also provide the R source package `hhh4contacts` containing the data and code to reproduce our analysis. Run `demo("hhh4contacts")` after installing and loading the package.

Funding

This work was supported by the Swiss National Science Foundation [project #137919].

Acknowledgements

We thank the associate editor, two anonymous referees, and Michael Höhle for helpful comments on a previous version of this manuscript. Joël Mossong made the POLYMOD data available.

References

Atmar, Robert L., Opekun, Antone R., Gilger, Mark A., Estes, Mary K., Crawford, Sue E., Neill, Frederick H. and Graham, David Y. (2008). Norwalk virus shedding after experimental human infection. *Emerging Infectious Diseases* **14**(10), 1553–1557.

Baguelin, M., Flasche, S., Camacho, A., Demiris, N., Miller, E. and Edmunds, W. J. (2013). Assessing optimal target populations for influenza vaccination programmes: An evidence synthesis and modelling study. *PLOS Medicine* **10**(10), e1001527.

Barclay, V. C., Smieszek, T., He, J., Cao, G., Rainey, J. J., Gao, H., Uzicanin, A. and Salathé, M. (2014). Positive network assortativity of influenza vaccination at a high school: Implications for outbreak risk and herd immunity. *PLOS ONE* **9**(2), e87042.

Bernard, H., Faber, M., Wilking, H., Haller, S., Höhle, M., Schielke, A., Ducomble, T., Siffczyk, C., Merbecks, S. S., Fricke, G., Hamouda, O., Stark, K., Werber, D. *and others*. (2014*a*). Large multistate outbreak of norovirus gastroenteritis associated with frozen strawberries, Germany, 2012. *Eurosurveillance* **19**(8), pii=20719.

Bernard, H., Höhne, M., Niendorf, S., Altmann, D. and Stark, K. (2014*b*). Epidemiology of norovirus gastroenteritis in Germany 2001–2009: Eight seasons of routine surveillance. *Epidemiology & Infection* **142**(1), 63–74.

Bernard, H., Werber, D. and Höhle, M. (2014*c*). Estimating the under-reporting of norovirus illness in Germany utilizing enhanced awareness of diarrhoea during a large outbreak of Shiga toxin-producing E. coli O104:H4 in 2011 – a time series analysis. *BMC Infectious Diseases* **14**(1), 116.

Birrell, P. J., Ketsetzis, G., Gay, N. J., Cooper, B. S., Presanis, A. M., Harris, R. J., Charlett, A., Zhang, X. S., White, P. J., Pebody, R. G. *and others*. (2011). Bayesian modeling to unmask and predict influenza A/H1N1pdm dynamics in London. *Proceedings of the National Academy of Sciences of the United States of America* **108**(45), 18238–18243.

Brockmann, D., Hufnagel, L. and Geisel, T. (2006). The scaling laws of human travel. *Nature* **439**(7075), 462–465.

Carrothers, G. A. P. (1956). An historical review of the gravity and potential concepts of human interaction. *Journal of the American Institute of Planners* **22**(2), 94–102.

Cliff, A. D. and Ord, J. K. (1975). Model building and the analysis of spatial pattern in human geography. *Journal of the Royal Statistical Society. Series B (Methodological)* **37**(3), 297–348.

Farrington, C. P., Kanaan, M. N. and Gay, N. J. (2003). Branching process models for surveillance of infectious diseases controlled by mass vaccination. *Biostatistics* **4**(2), 279–295.

Geilhufe, M., Held, L., Skrøvseth, S. O., Simonsen, G. S. and Godtliebsen, F. (2014). Power law approximations of movement network data for modeling infectious disease spread. *Biometrical Journal* **56**(3), 363–382.

Gibbons, C. L., Mangen, M.-J., Plass, D., Havelaar, A. H., Brooke, R. J., Kramarz, P., Peterson, K. L., Stuurman, A. L., Cassini, A., Fèvre, E. M. *and others.* (2014). Measuring underreporting and under-ascertainment in infectious disease datasets: a comparison of methods. *BMC Public Health* **14**(1), 1–17.

Goeyvaerts, N., Hens, N., Ogunjimi, B., Aerts, M., Shkedy, Z., van Damme, P. and Beutels, P. (2010). Estimating infectious disease parameters from data on social contacts and serological status. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **59**(2), 255–277.

Goeyvaerts, N., Willem, L., van Kerckhove, K., Vandendijck, Y., Hanquet, G., Beutels, P. and Hens, N. (2015). Estimating dynamic transmission model parameters for seasonal influenza by fitting to age and season-specific influenza-like illness incidence. *Epidemics* **13**, 1–9.

Graham, David Y., Jiang, Xi, Tanaka, Tomoyuki, Opekun, Antone R., Madore, H. Paul and Estes, Mary K. (1994). Norwalk virus infection of volunteers: new insights based on improved assays. *Journal of Infectious Diseases* **170**(1), 34–43.

Grimmett, G. R. and Stirzaker, D. R. (2001). *Probability and Random Processes*, 3rd edition. Oxford: Oxford University Press.

Götz, H., Ekdahl, K., Lindbäck, J., de Jong, B., Hedlund, K. O. and Giesecke, J. (2001). Clinical spectrum and transmission characteristics of infection with Norwalk-like virus: Findings from a large community outbreak in Sweden. *Clinical Infectious Diseases* **33**(5), 622–628.

Heijne, J. C. M., Teunis, P., Morroy, G., Wijkmans, C., Oostveen, S., Duizer, E., Kretzschmar, M. and Wallinga, J. (2009). Enhanced hygiene measures and norovirus transmission during an outbreak. *Emerging Infectious Diseases* **15**(1), 24–30.

Held, L., Höhle, M. and Hofmann, M. (2005). A statistical framework for the analysis of multivariate infectious disease surveillance counts. *Statistical Modelling* **5**(3), 187–199.

Held, L. and Paul, M. (2012). Modeling seasonality in space-time infectious disease surveillance data. *Biometrical Journal* **54**(6), 824–843.

Held, L. and Sabanés Bové, D. (2014). *Applied Statistical Inference: Likelihood and Bayes*. Berlin: Springer.

Hens, N., Ayele, G., Goeyvaerts, N., Aerts, M., Mossong, J., Edmunds, J. and Beutels, P. (2009). Estimating the impact of school closure on social mixing behaviour and the transmission of close contact infections in eight European countries. *BMC Infectious Diseases* **9**(1), 187.

Herzog, S. A., Paul, M. and Held, L. (2011). Heterogeneity in vaccination coverage explains the size and occurrence of measles epidemics in German surveillance data. *Epidemiology & Infection* **139**(04), 505–515.

Höhle, M. (2016). Infectious Disease Modelling. In: Lawson, A. B., Banerjee, S., Haining, R. P. and Ugarte, M. D. (editors), *Handbook of Spatial Epidemiology*, Chapman & Hall/CRC Handbooks of Modern Statistical Methods. Chapman and Hall/CRC. Forthcoming.

Kucharski, A. J., Conlan, A. J. K. and Eames, K. T. D. (2015). School's out: seasonal variation in the movement patterns of school children. *PLOS ONE* **10**(6), 1–10.

Küchenhoff, H., Mwalili, S. M. and Lesaffre, E. (2006). A general method for dealing with misclassification in regression: The misclassification SIMEX. *Biometrics* **62**(1), 85–96.

Marshall, J. A. and Bruggink, L. D. (2011). The dynamics of norovirus outbreak epidemics: Recent insights. *International Journal of Environmental Research and Public Health* **8**(4), 1141–1149.

Meyer, S. and Held, L. (2014). Power-law models for infectious disease spread. *Annals of Applied Statistics* **8**(3), 1612–1639.

Meyer, S., Held, L. and Höhle, M. (2016). Spatio-temporal analysis of epidemic phenomena using the R package `surveillance`. *Journal of Statistical Software*. In press. Preprint available from `http://arxiv.org/abs/1411.0416`.

Mossong, J., Hens, N., Jit, M., Beutels, P., Auranen, K., Mikolajczyk, R., Massari, M., Salmaso, S., Tomba, G. S., Wallinga, J., Heijne, J., Sadkowska-Todys, M., Rosinska, M. *and others.* (2008). Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Medicine* **5**(3), e74.

O'Dea, E. B., Pepin, K. M., Lopman, B. A. and Wilke, C. O. (2014). Fitting outbreak models to data from many small norovirus outbreaks. *Epidemics* **6**, 18–29.

Paul, M. and Held, L. (2011). Predictive assessment of a non-linear random effects model for multivariate time series of infectious disease counts. *Statistics in Medicine* **30**(10), 1118–1136.

Paul, M., Held, L. and Toschke, A. (2008). Multivariate modelling of infectious disease surveillance data. *Statistics in Medicine* **27**(29), 6250–6267.

Pringle, K., Lopman, B., Vega, E., Vinje, J., Parashar, U. D. and Hall, A. J. (2015). Noroviruses: Epidemiology, immunity and prospects for prevention. *Future Microbiology* **10**(1), 53–67.

R Core Team. (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Read, J. M., Edmunds, W. J., Riley, S., Lessler, J. and Cummings, D. A. T. (2012). Close encounters of the infectious kind: Methods to measure social mixing behaviour. *Epidemiology & Infection* **140**(12), 2117–2130.

Read, J. M., Lessler, J., Riley, S., Wang, S., Tan, L. J., Kwok, K. O., Guan, Y., Jiang, C. Q. and Cummings, D. A. T. (2014). Social mixing patterns in rural and urban areas of southern China. *Proceedings of the Royal Society of London B: Biological Sciences* **281**(1785), 20140268.

Robert Koch-Institut. (2010). Überarbeitete Falldefinitionen zur Übermittlung von Nachweisen von Denguevirus und Norovirus und Erkrankungs- oder Todesfällen an Denguefieber und an Norovirus-Gastroenteritis. *Epidemiologisches Bulletin* **2010**(49), 494–495. In German.

Rohani, P., Zhong, X. and King, A. A. (2010). Contact network structure explains the changing epidemiology of pertussis. *Science* **330**(6006), 982–985.

Schrödle, B., Held, L. and Rue, H. (2012). Assessing the impact of a movement network on the spatiotemporal spread of infectious diseases. *Biometrics* **68**(3), 736–744.

Smieszek, Timo, Barclay, Victoria, Seeni, Indulaxmi, Rainey, Jeanette, Gao, Hongjiang, Uzicanin, Amra and Salathé, Marcel. (2014). How should social mixing be measured: comparing web-based survey and sensor-based methods. *BMC Infectious Diseases* **14**(1), 136.

Sukhrie, Faizel H. A., Teunis, Peter, Vennema, Harry, Copra, Cedrick, Thijs Beersma, Matthias F. C., Bogerman, Jolanda and Koopmans, Marion. (2012). Nosocomial transmission of norovirus is mainly caused by symptomatic cases. *Clinical Infectious Diseases* **54**(7), 931–937.

van Asten, L., van den Wijngaard, C., van Pelt, W., van de Kassteele, J., Meijer, A., van der Hoek, W., Kretzschmar, M. and Koopmans, M. (2012). Mortality attributable to 9 common infections: Significant effect of influenza A, respiratory syncytial virus, influenza B, norovirus, and parainfluenza in elderly persons. *Journal of Infectious Diseases* **206**(5), 628–639.

van Kerckhove, K., Hens, N., Edmunds, W. J. and Eames, K. T. D. (2013). The impact of illness on social networks: Implications for transmission and control of influenza. *American Journal of Epidemiology* **178**(11), 1655–1662.

Wakefield, J. (2007). Disease mapping and spatial regression with count data. *Biostatistics* **8**(2), 158–183.

Wallinga, J., Edmunds, W. J. and Kretzschmar, M. (1999). Perspective: human contact patterns and the spread of airborne infectious diseases. *Trends in Microbiology* **7**(9), 372–377.

Wallinga, J., Teunis, P. and Kretzschmar, M. (2006). Using data on social contacts to estimate age-specific transmission parameters for respiratory-spread infectious agents. *American Journal of Epidemiology* **164**(10), 936–944.

Worby, C. J., Chaves, S. S., Wallinga, J., Lipsitch, M., Finelli, L. and Goldstein, E. (2015). On the relative role of different age groups in influenza epidemics. *Epidemics* **13**, 10–16.

Xia, Y., Bjørnstad, O. N. and Grenfell, B. T. (2004). Measles metapopulation dynamics: A gravity model for epidemiological coupling and dynamics. *The American Naturalist* **164**(2), 267–281.

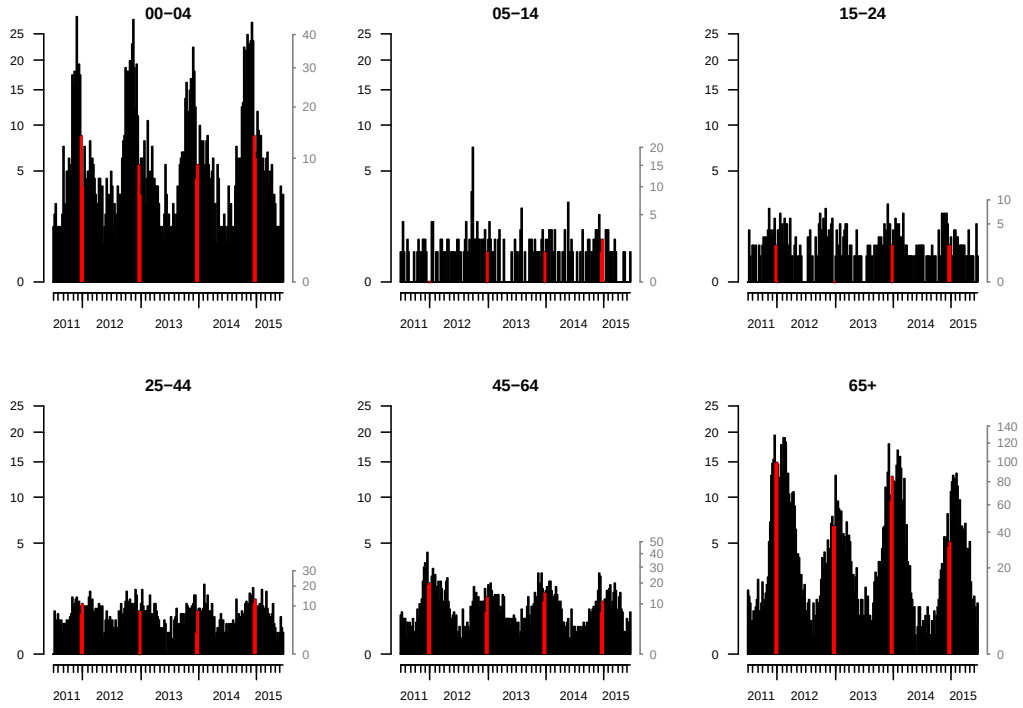


Fig. 1: Weekly age-stratified norovirus gastroenteritis incidence (per 100 000 inhabitants) in Berlin. The incidence on the left axis obeys the same $\sqrt{}$-scale in all panels, while the corresponding counts can be read off from the right axis. The population sizes of the age groups are 160 885, 271 574, 365 536, 1 064 326, 976 284 and 663 267, respectively. The yearly Christmas break in calendar weeks 52 and 1 is highlighted.

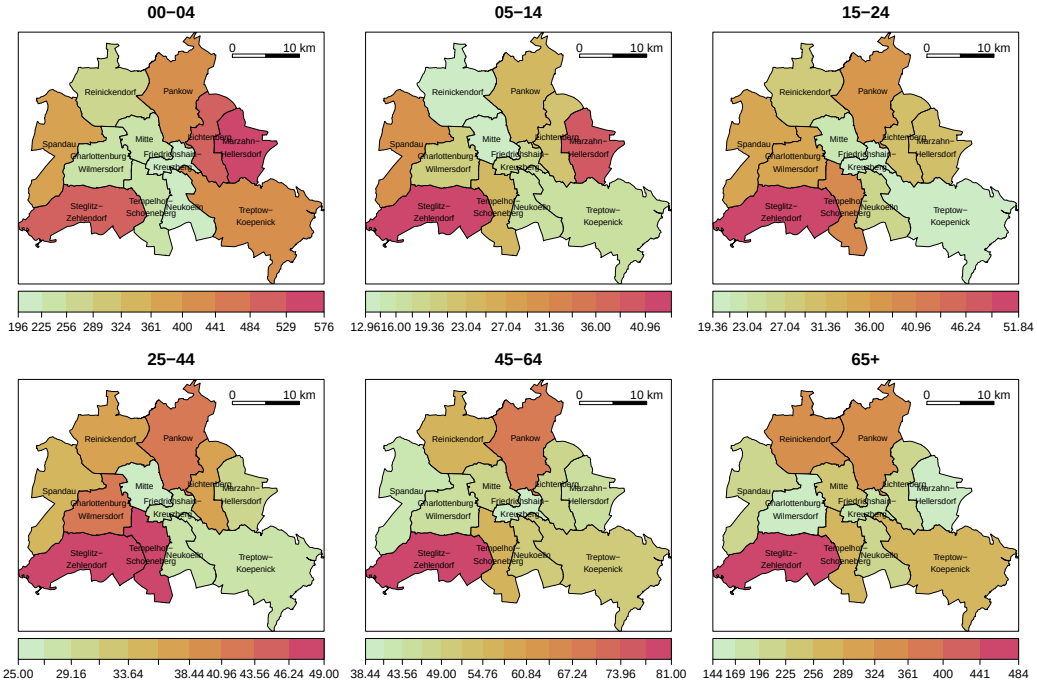


Fig. 2: Age-stratified maps of mean yearly norovirus gastroenteritis incidence (per 100 000 inhabitants) in Berlin's city districts, from 2011-W27 to 2015-W26.

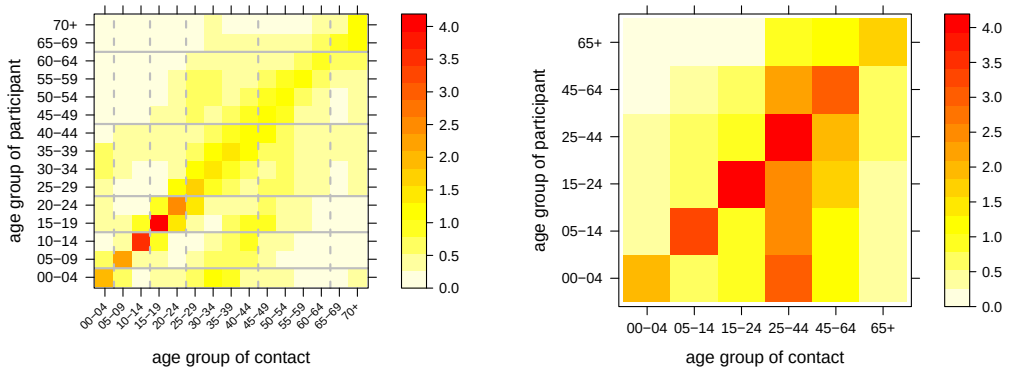
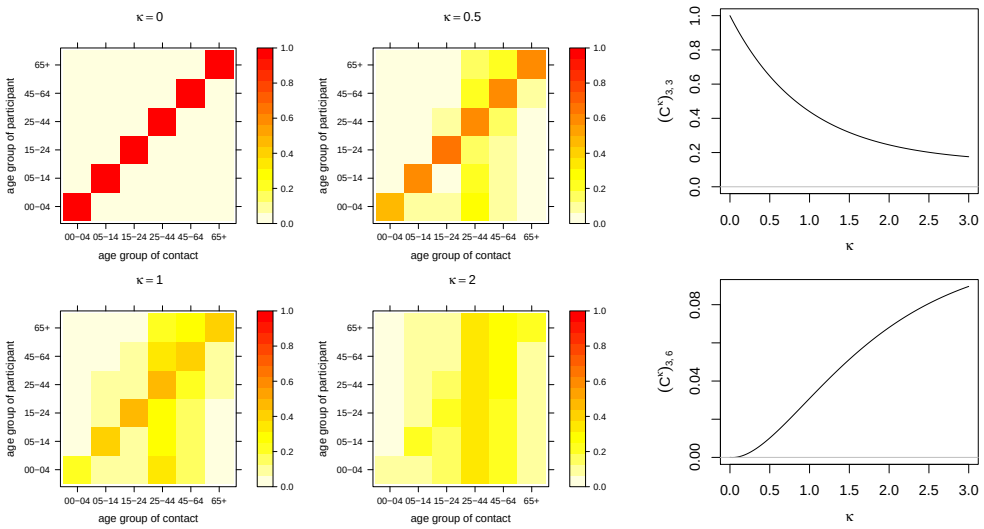


Fig. 3: Age-structured contact matrix estimated from the German POLYMOD sample with 5-year intervals (left) and its aggregated version with age groups matching the surveillance data (right). The entries refer to the mean number of contact persons per participant per day.



(a) $\boldsymbol{C}^\kappa$ for different values of κ. Moving from the original matrix at $\kappa = 1$ to the limiting case $\kappa = 0$, off-diagonal values are pushed row-wise towards the diagonal. For increasing κ the row-wise distributions approach each other.

(b) The diagonal entry $[3, 3]$ and the off-diagonal entry $[3, 6]$ of $\boldsymbol{C}^\kappa$ as a function of κ.

Fig. 4: The effect of raising the row-normalized version of the aggregated contact matrix $\boldsymbol{C}$ from Figure 3 to the power κ as defined in (3.5).

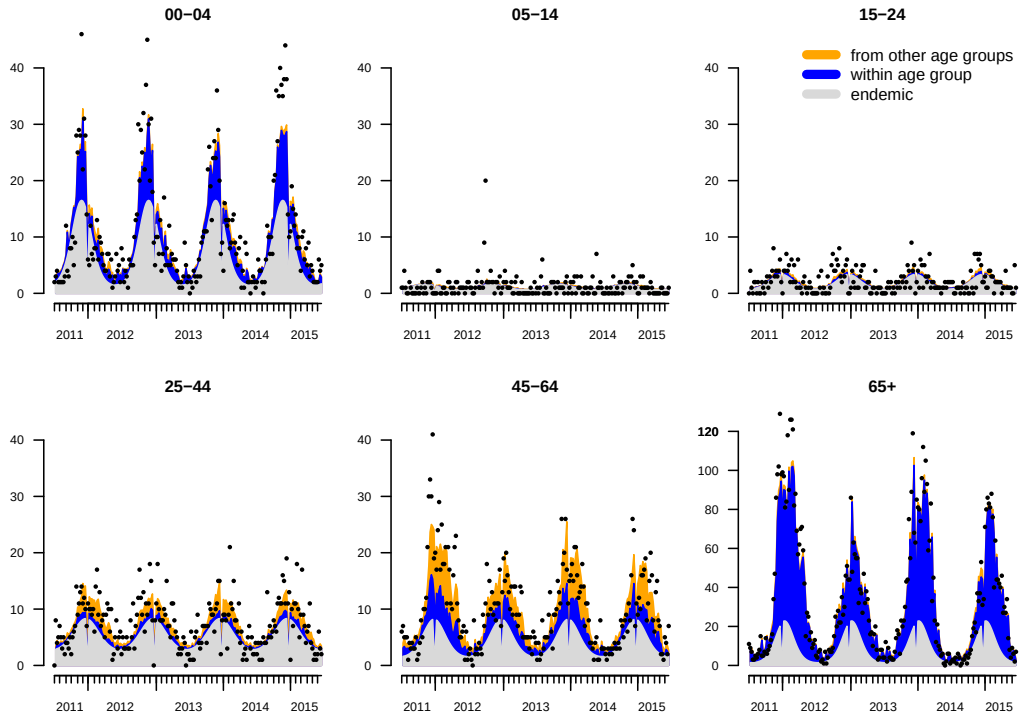


Fig. 5: Fitted mean components from the AIC-optimal model, aggregated over all districts. The dots correspond to the reported numbers of cases.

Table 1: Model summaries for the age-stratified, areal surveillance data of norovirus gastroenteritis in Berlin. For reference, the first row represents the purely endemic model, which assumes independent counts. The remaining rows correspond to endemic-epidemic models with a spatial power law but varying assumptions on the contact matrix $\boldsymbol{C}$ between age groups. The columns refer to the following model characteristics: the number of parameters, the difference in Akaike's Information Criterion compared to the purely endemic model, the power τ of the population scaling factor, the decay parameter ρ of the spatial power law, and the power adjustment κ of the contact matrix. The parameter columns contain the estimates and 95% Wald confidence intervals.

	dim	ΔAIC	τ	ρ	κ
purely endemic model	36	0.0	–	–	–
homogeneous mixing ($\boldsymbol{C} = \mathbf{1}$)	55	-415.4	1.19 (0.83 to 1.55)	2.43 (2.04 to 2.88)	–
no mixing ($\boldsymbol{C} = \boldsymbol{I}$)	55	-602.8	0.61 (0.24 to 0.98)	2.18 (1.89 to 2.53)	–
original contact matrix $\boldsymbol{C}$	55	-631.9	0.97 (0.66 to 1.28)	2.34 (2.03 to 2.70)	–
adjusted contact matrix $\boldsymbol{C}^\kappa$	56	-659.4	0.86 (0.53 to 1.19)	2.27 (1.98 to 2.61)	0.47 (0.34 to 0.66)
based on physical contacts only	56	-655.3	0.85 (0.52 to 1.19)	2.27 (1.98 to 2.61)	0.48 (0.35 to 0.66)

Incorporating social contact data in spatio-temporal models for infectious disease spread

Supplementary Materials

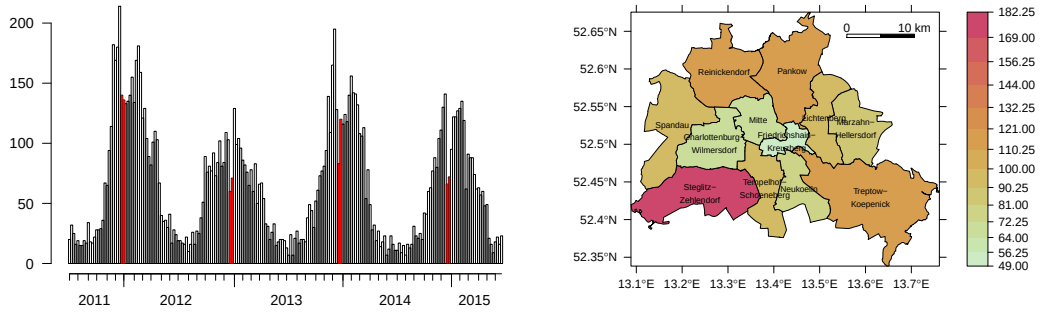
SEBASTIAN MEYER*, LEONHARD HELD

Epidemiology, Biostatistics and Prevention Institute, University of Zurich, Hirschengraben 84, CH-8001 Zürich, Switzerland

sebastian.meyer@uzh.ch

This document contains additional figures from our analysis of norovirus infections by age group in Berlin, 2011–2015. Further supplementary files contain an animation of the data and the R source package `hhh4contacts`, respectively.

*To whom correspondence should be addressed.



(a) Count time series. The yearly Christmas break in calendar weeks 52 and 1 is highlighted.

(b) Mean yearly incidence by city district (per 100 000 inhabitants).

Fig. S1: Norovirus gastroenteritis in Berlin, from 2011-W27 to 2015-W26.

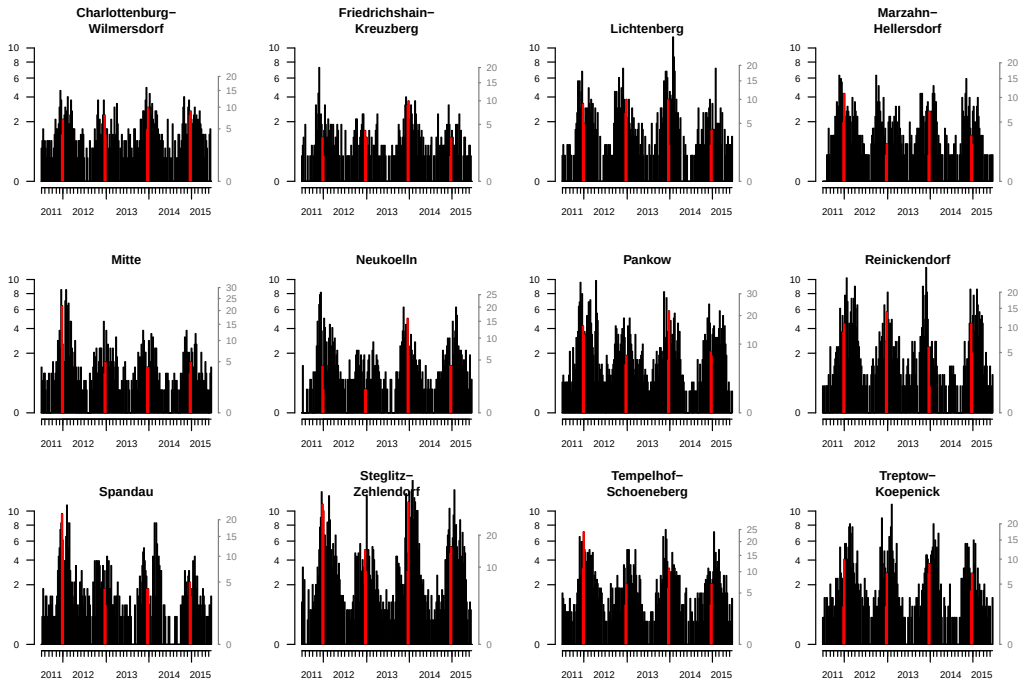


Fig. S2: Weekly norovirus gastroenteritis incidence (per 100 000 inhabitants) stratified by Berlin's 12 city districts. The incidence on the left axis obeys the same $\sqrt{}$-scale in all panels, while the corresponding counts can be read off from the right axis. The yearly Christmas break in calendar weeks 52 and 1 is highlighted.

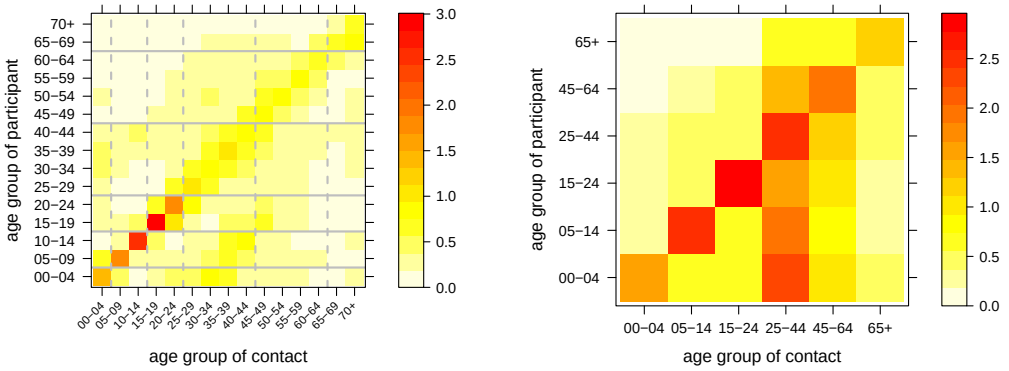


Fig. S3: Age-structured *physical* contact matrix estimated from the German POLYMOD sample with 5-year intervals (left) and its aggregated version with age groups matching the surveillance data (right). The entries refer to the mean number of persons contacted physically per participant per day.

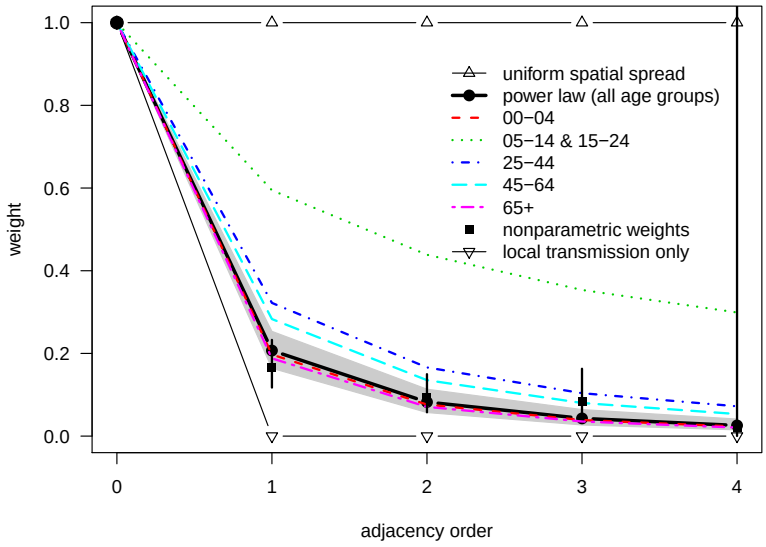


Fig. S4: Estimated (group-specific) power-law decays as well as unconstrained estimates of the transmission weights $w_{r'r}$ as a function of adjacency order. The grey shading represents a 95% confidence interval for the joint power law, as do the vertical bars on top of the unconstrained estimates.

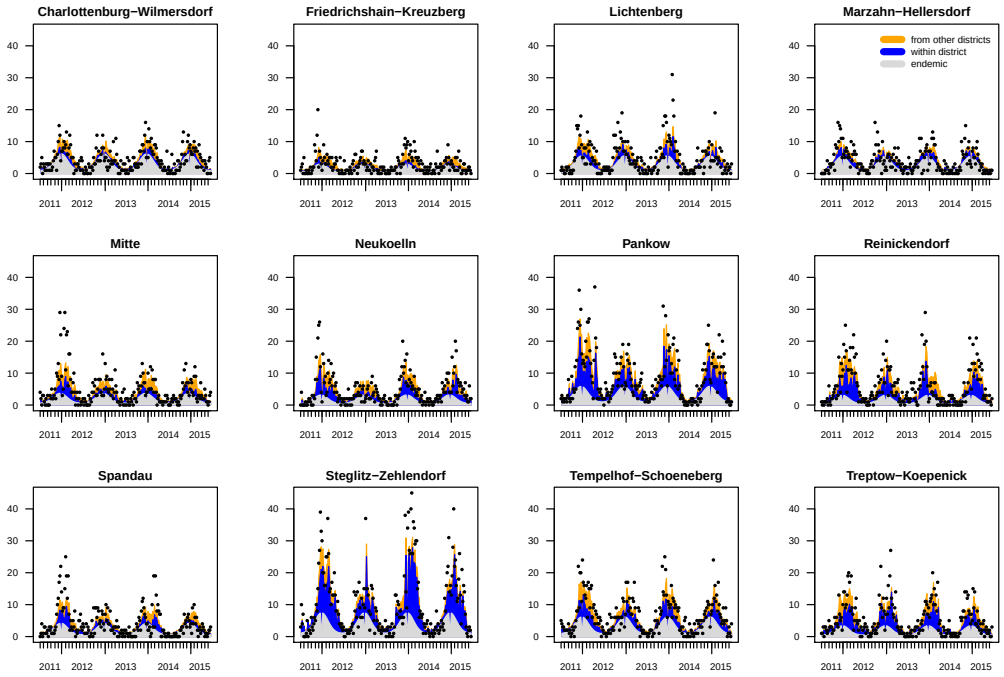


Fig. S5: Fitted mean components by district aggregated over all age groups. The dots correspond to the reported numbers of cases.

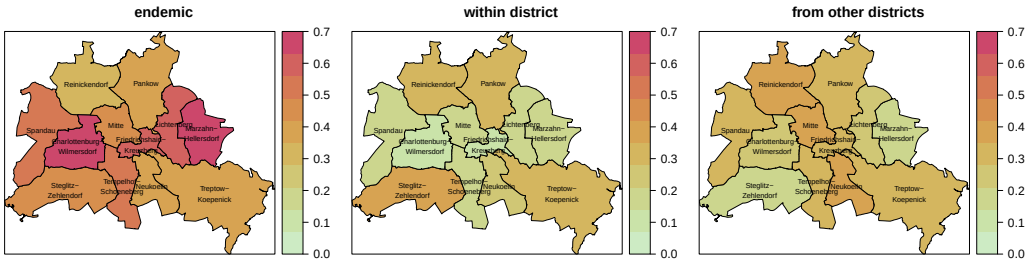


Fig. S6: These maps show the proportions of the district-specific mean attributed to various model components (for each district, the values of the three maps sum to 1). For this purpose, the fitted mean $\hat{\mu}_{grt}$ is aggregated over age groups and averaged over weeks.

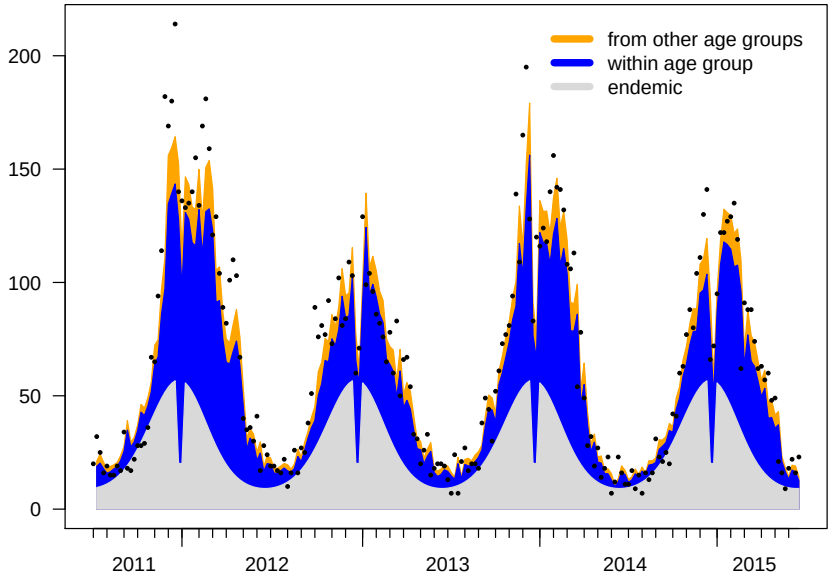


Fig. S7: Fitted mean components from the AIC-optimal model, aggregated over all districts and all age groups. The dots correspond to the reported numbers of cases.

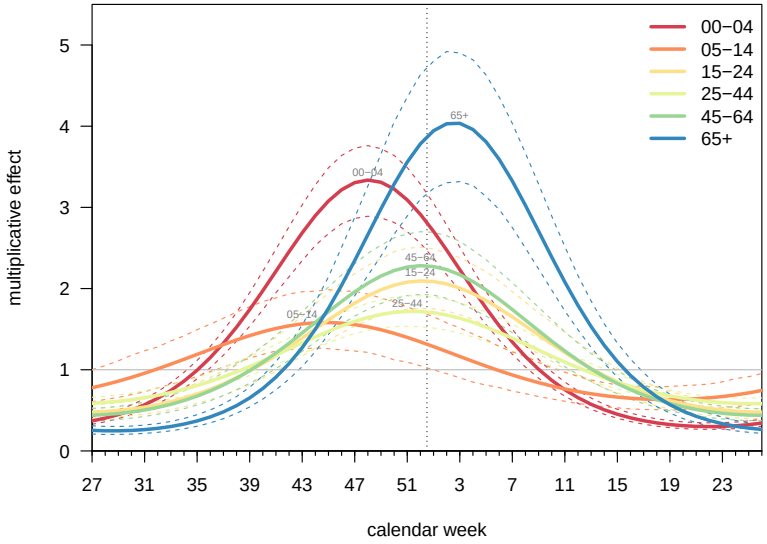
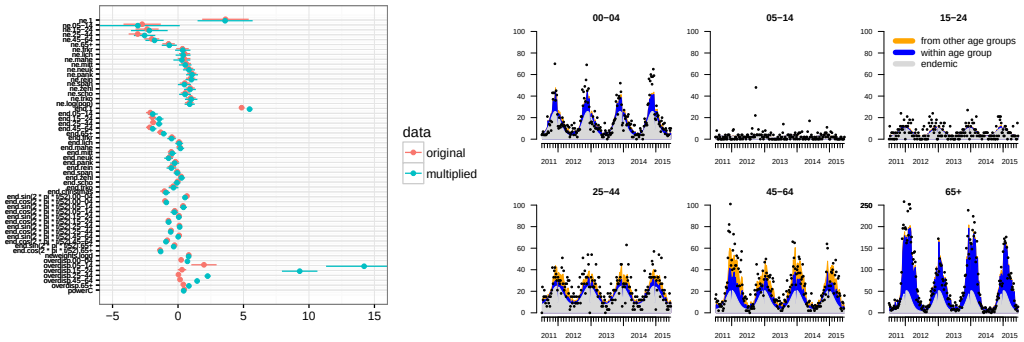


Fig. S8: Estimated age-dependent seasonality of the endemic model component with 95% point-wise confidence intervals.



(a) Comparison of parameter estimates.

(b) Fitted mean components.

Fig. S9: AIC-optimal model fitted to counts multiplied by age-specific under-reporting factors.

PAPER IV

Model-based testing for space-time interaction using point processes: An application to psychiatric hospital admissions in an urban area

Sebastian Meyer, Ingeborg Warnke, Wulf Rössler, Leonhard Held

Published in *Spatial and Spatio-temporal Epidemiology*, 2016, **17**, 15–25.

Spatial and Spatio-temporal Epidemiology 17 (2016) 15–25

Contents lists available at ScienceDirect

Spatial and Spatio-temporal Epidemiology

journal homepage: www.elsevier.com/locate/sste

Original Research

Model-based testing for space–time interaction using point processes: An application to psychiatric hospital admissions in an urban area

Sebastian Meyer[a,*], Ingeborg Warnke[b], Wulf Rössler[c], Leonhard Held[a]

[a] Epidemiology, Biostatistics and Prevention Institute (EBPI), University of Zurich, Hirschengraben 84, 8001 Zürich, Switzerland
[b] Department of Psychiatry, Psychotherapy and Psychosomatics (DPPP), University Hospital of Psychiatry, Lenggstrasse 31, 8032 Zürich, Switzerland
[c] University Hospital of Psychiatry, Militärstrasse 8, 8021 Zurich, Switzerland

ARTICLE INFO

Article history:
Received 13 August 2015
Revised 26 February 2016
Accepted 21 March 2016
Available online 13 April 2016

Keywords:
Spatio-temporal point process
Knox test
Mantel test
Space–time *K*-function
Global test of clustering
Psychiatric inpatient admissions

ABSTRACT

Spatio-temporal interaction is inherent to cases of infectious diseases and occurrences of earthquakes, whereas the spread of other events, such as cancer or crime, is less evident. Statistical significance tests of space–time clustering usually assess the correlation between the spatial and temporal (transformed) distances of the events. Although appealing through simplicity, these classical tests do not adjust for the underlying population nor can they account for a distance decay of interaction. We propose to use the framework of an endemic–epidemic point process model to jointly estimate a background event rate explained by seasonal and areal characteristics, as well as a superposed epidemic component representing the hypothesis of interest. We illustrate this new model-based test for space–time interaction by analysing psychiatric inpatient admissions in Zurich, Switzerland (2007–2012). Several socio-economic factors were found to be associated with the admission rate, but there was no evidence of general clustering of the cases.

1. Introduction

Cases of infectious diseases naturally exhibit spatio-temporal interaction. Once an individual becomes infected, it may cause secondary cases by transmitting the infectious agent to susceptible individuals. If the force of infection triggered by an infectious individual decreases with time and distance, cases are likely to appear in spatio-temporal clusters, i.e., cases which are close in time tend to also be close in space. This contagion goes beyond any purely spatial clustering in densely populated areas or seasonal forcing. Spatio-temporal clustering can be observed in other fields as well, for example, seismology (Ogata, 1998), veterinary epidemiology (Ward and Carpenter, 2000), cancer epidemiology (Birch et al., 2000; Gustafsson and Carstensen, 2000), criminology (Grubesic and Mack, 2008; Johnson, 2010), and transportation research (Eckley and Curtin, 2013).

The spread of various mental health indicators in local social networks has been analysed: suicide (Haw et al., 2013), happiness (Fowler and Christakis, 2008), depression (Rosenquist et al., 2011), and even autism (Liu et al., 2010). Usually, this spread occurs within a short time period and in geographically close living areas. For hospital admissions of patients with mental disorders, previous studies have found associations with local socio-economic factors such as unemployment rate, average income, proportion of one-person households, proportion of foreigners, or level of

* Corresponding author.
E-mail addresses: sebastian.meyer@uzh.ch (S. Meyer), leonhard.held@uzh.ch (L. Held).

http://dx.doi.org/10.1016/j.sste.2016.03.002
1877-5845/

urbanisation (see, e.g., Simone et al., 2013). However, to the best of our knowledge, space–time clustering of psychiatric hospital admissions has not been investigated so far, i.e., if admissions close in time are also close in space with respect to where the patients live. The idea is similar to the spread of rumours (Daley and Gani, 1999, Chapter 5), where here the rumour is the news that someone in the neighbourhood searches professional psychiatric treatment. For mentally stricken citizens, such news could lower help-seeking barriers, whereas people with a treatment history might remember own positive experiences by the knowledge about others' hospital admissions.

The first statistical significance test for "low intensity epidemicity" is due to Knox (1964; 1963), who investigated space–time interaction of cases of cleft lip and palate. Several variations of his test have since been developed, for example, by generalizing the test statistic as a measure of correlation of spatial and temporal closeness (Mantel, 1967), or by replacing Euclidean distances by k nearest neighbours (Jacquez, 1996). An enhanced version of the Knox statistic derived from point process theory is the spatio-temporal K-function by Diggle et al. (1995). This approach enables a formal definition of the null hypothesis of no space–time interaction as a process with independent spatial and temporal components. However, these classical tests do not account for spatial or temporal inhomogeneity, e.g., variations in population density or seasonal effects, and are thus prone to bias (Kulldorff and Hjalmars, 1999; Mack et al., 2012).

Second-order characteristics of spatio-temporal point processes, such as the K-function, have recently been generalized to the inhomogeneous case (Gabriel and Diggle, 2009; Møller and Ghorbani, 2012). In practice, adjusting the K-function for spatial or temporal inhomogeneity requires a two-step procedure. The crucial first step is to obtain estimates of the spatial and temporal intensity functions. As in "empirical" point process models (Diggle, 2013, Chapter 12; Diggle et al., 2005), non-parametric kernel methods are commonly employed. Supplied with area-level data on relevant socio-economic factors, we will instead analyse interaction of events using a *mechanistic* spatio-temporal point process model. Such models are especially attractive if covariates can explain part of the heterogeneity in the observed point pattern, or if there is an explicit formulation of how past events affect the evolution of the process. For instance, mechanistic models have recently been applied to a range of epidemic phenomena, including residential burglaries (Mohler et al., 2011), infectious disease occurrence (Meyer et al., 2012), invasive plant species (Balderama et al., 2012), and epidemics among dolphins (Morris et al., 2015).

We propose to embed the test for space–time interaction of events in the *endemic–epidemic* point process regression framework of Meyer et al. (2012), which is implemented in the open-source R package surveillance (Meyer et al., 2016). The basic model formulation borrows from both the Epidemic-Type Aftershock-Sequences (ETAS) model (Ogata, 1998) and the multivariate time-series model for infectious disease counts proposed by Held et al. (2005). While the endemic model component reflects background heterogeneity due to, e.g., population structure and seasonality, the epidemic component makes the process "self-exciting" and causes spatio-temporal interaction. The basic idea of the proposed model-based test is to assess the evidence for an epidemic component with a Monte Carlo permutation approach (Besag and Diggle, 1977).

This paper is organized as follows: Section 2 briefly reviews the two most popular classical tests for space–time interaction, the Knox and the Mantel tests, as well as the space–time K-function. The new test procedure via an endemic–epidemic point process model is introduced in Section 3. In Section 4, we apply the various tests to psychiatric hospital admissions in the city of Zurich (Switzerland), 2007–2012, and the supplementary material shows corresponding results for the invasive meningococcal disease data originally analysed by Meyer et al. (2012). Section 5 concludes the paper and Appendix A lists the software used to perform the analyses.

2. Classical approaches to testing for space–time interaction

Given a point pattern $\{(\boldsymbol{s}_i, t_i) : i = 1, \ldots, n\}$ with spatial coordinates $\boldsymbol{s}_i$ and time points t_i observed in a region $\boldsymbol{W}$ during a period $(0, T]$, classical tests for space–time interaction basically check if events closer in time also tend to be closer in space. How closeness is measured varies between the different tests, but many use the form of test statistic put forward by Mantel (1967),

$$T_0 \propto \sum_{i=1}^{n} \sum_{j \neq i} a_{ij}^{\mathrm{s}} a_{ij}^{\mathrm{t}}\,, \tag{1}$$

where a_{ij}^{s} and a_{ij}^{t} are measures of the spatial and temporal adjacency of the events i and j, respectively (Marshall, 1991). These measures are often defined as a function of the respective Euclidean distances $d_{ij}^{\mathrm{s}} = \|\boldsymbol{s}_i - \boldsymbol{s}_j\|$ and $d_{ij}^{\mathrm{t}} = |t_i - t_j|$. Larger values of T_0 support the (one-sided) alternative hypothesis of positive association between spatial and temporal adjacency.

The distribution of T_0 under the null hypothesis of no space–time interaction is determined by a Monte Carlo permutation approach (Besag and Diggle, 1977). In each of $B = 999$, say, replications, the test statistic is computed for a random permutation of the time labels while holding the locations fixed. This destroys any systematic space–time interaction in the data but leaves both the marginal spatial and temporal distributions unchanged. The (one-sided) p-value for a positive association between spatial and temporal adjacency (i.e., clustering) is then obtained as the proportion of test statistics greater than or equal to the observed one. The smallest attainable p-value from this Monte Carlo procedure thus is $1/(B+1) = 0.001$ (the "resolution").

In the following subsections, we outline the various choices of adjacency measures employed by three different tests. For a broader overview of space–time interaction tests and references to applications we point to Tango (2010, Chapter 7) and Mack et al. (2012, Section 2).

2.1. Knox test

For the Knox test, critical distances in space (δ) and time (τ) have to be specified to yield a categorization of the distances into "close" vs. "not close". The test statistic is then defined as the number of event pairs close both in space and time according to these distance thresholds:

$$T_{\text{Knox}} = \frac{1}{2}\sum_{i=1}^{n}\sum_{j\neq i}\mathbb{1}(d_{ij}^{s}\leq\delta)\mathbb{1}(d_{ij}^{t}\leq\tau)\,. \tag{2}$$

If the point pattern exhibits clustering at the predefined spatio-temporal scales, the observed number of close pairs will be larger than the expected number under the null hypothesis of no space–time interaction.

The Knox test is appealing through simplicity but criticized for the subjectivity in specifying δ and τ. It is thus often applied over a range of critical distances (Grubesic and Mack, 2008).

2.2. Mantel test

Mantel (1967) elaborates on the general use of the test statistic (1) with any suitable distance measures for the application at hand. This includes the indicator functions of the Knox test, raw or transformed Euclidean distances such as $1/(d_{ij}+1)$ (to collapse the range of large distances), but also other distance measures like travel time. An often used standardized version of Mantel's test statistic is the Pearson correlation between the spatial and temporal distances of all event pairs, i.e.,

$$T_{\text{Mantel}} = \frac{1}{n(n-1)-2}\sum_{i=1}^{n}\sum_{j\neq i}\frac{d_{ij}^{s}-\bar{d}_{s}}{\hat{\sigma}_{d^{s}}}\frac{d_{ij}^{t}-\bar{d}_{t}}{\hat{\sigma}_{d^{t}}}\,, \tag{3}$$

where the $\bar{d}_{\cdot}$ and $\hat{\sigma}_{d^{\cdot}}$ symbols denote the respective sample means and standard deviations of the $n(n-1)/2$ pairwise distances. Note that the standardized Mantel test statistic formulated in Jacquez (1996, Section 1.2), Ward and Carpenter (2000, Section 2.1), and Mack et al. (2012, Section 2.2) does *not* exactly correspond to the Pearson correlation (3) as used in this paper and in many open-source software packages implementing the Mantel test.[1]

Unlike the Knox test, the standardized Mantel test does not require the specification of distance thresholds. However, it assesses a linear relationship between spatial and temporal distances, which might not be appropriate over the whole distance range (the clustering is expected to occur towards the origin in a plot of spatial against temporal distance). The reciprocal transformation of the pairwise distances suggested by Mantel (1967) would then again depend on the choice of a suitable constant to avoid division by zero.

2.3. Space–time K-function analysis

The space–time K-function $K(\delta,\tau)$ essentially interprets the Knox statistic (2) as a function of the critical distances δ and τ while accounting for edge effects (Diggle et al., 1995). For stationary point processes, the K-function is proportional to the expected number of further events occurring within distance δ and time τ of an arbitrary event (Diggle, 2013, Chapters 10 and 11). An approximately unbiased estimator of the K-function is

$$\hat{K}(\delta,\tau) = \frac{|W|T}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}w_{ij}v_{ij}\mathbb{1}(d_{ij}^{s}\leq\delta)\mathbb{1}(d_{ij}^{t}\leq\tau)\,, \tag{4}$$

where w_{ij} and v_{ij} are edge-correction weights (Diggle et al., 1995, Section 3). Note that this estimator is proportional to the Knox statistic (2) if the weights all equal unity. Similar estimators $\hat{K}_{s}(\delta)$ and $\hat{K}_{t}(\tau)$ exist for the purely spatial and temporal component processes, respectively. The test statistic is then derived from the property that the spatio-temporal K-function factorizes into the component K-functions under the null hypothesis of no space–time interaction. Specifically,

$$\hat{D}(\delta,\tau) = \hat{K}(\delta,\tau) - \hat{K}_{s}(\delta)\,\hat{K}_{t}(\tau) \tag{5}$$

has an expectation of zero for any given thresholds δ and τ. Diggle et al. (1995) recommend a perspective plot of the $\hat{D}$ surface to gain more insight into the relevant scales of spatio-temporal interaction than the previous tests. An omnibus test over a set of spatial and temporal scales is obtained by

$$T_{\text{Diggle}} = \sum_{\delta}\sum_{\tau}\hat{D}(\delta,\tau) \tag{6}$$

and evaluating statistical significance by the common Monte Carlo permutation approach.

This test is used far less frequently than the Knox or Mantel tests, since a (polygonal) representation of the observation region $\boldsymbol{W}$ is required in addition to the event times and locations to compute the edge-correction weights, which also needs specialized software. Some studies such as McNally et al. (2006) have used a simplification of this test procedure towards an omnibus Knox test by ignoring the edge-correction weights.

3. Model-based assessment of space–time interaction

The classical Knox and Mantel tests have the advantage of working solely on the observed event times and locations without the need for additional data. In practice, however, the question of space–time interaction is often accompanied by questions of how the event rate varies over time or space, how far the events interact, or if specific events trigger a higher amount of clustering. To address these issues, we propose to embed the test for space–time interaction in a regression framework for spatio-temporal point patterns of epidemic phenomena (Meyer et al., 2012; Meyer and Held, 2014).

3.1. Spatio-temporal point process model

The considered spatio-temporal point process model describes the conditional intensity $\lambda(\boldsymbol{s}, t)$ for an event at location $\boldsymbol{s}$ and time t (given the history of the

[1] We have inspected the R packages ecodist, vegan, and ade4 (all available at http://CRAN.R-project.org/), as well as the python packages PySAL (http://pysal.org/) and scikit-bio (http://github.com/biocore/scikit-bio/).

process) by a superposition of an endemic and an epidemic component:

$$\lambda(\boldsymbol{s},t) = \rho_{[\boldsymbol{s}][t]}\,\nu_{[\boldsymbol{s}][t]} + \sum_{j\in I(\boldsymbol{s},t)} \eta_j\, f(\|\boldsymbol{s}-\boldsymbol{s}_j\|)\, g(t-t_j)\,, \quad (7)$$

$$I(\boldsymbol{s},t) = \left\{ j : t_j < t \;\wedge\; t - t_j \le \tau_j \;\wedge\; \|\boldsymbol{s}-\boldsymbol{s}_j\| \le \delta_j \right\}.$$

The endemic component reflects the background rate of new events (cases) as explained by the offset $\rho_{[\boldsymbol{s}][t]}$, usually population density, and effects of other local and/or time-varying characteristics in the log-linear predictor $\nu_{[\boldsymbol{s}][t]}$. Here the spatial and temporal indices $[\boldsymbol{s}][t]$ refer to the regions and periods across which the covariates are collected, e.g., a district $\times$ week grid. Interpretation of the stand-alone endemic model component is thus straightforward since it is equivalent to a Poisson regression model for the aggregated number of events across the cells of the chosen grid (Meyer et al., 2016, Section 3.1). Note that this data-driven formulation leads to a piecewise constant endemic intensity, which – depending on the chosen grid and the underlying event-generating process – may not be a maintainable simplification.

The second, observation-driven epidemic component adds "infection pressure" from the set $I(\boldsymbol{s}, t)$ of past events and thus causes spatio-temporal interaction. During its infectious period of length τ_j and within its spatial interaction radius δ_j, the model assumes each event j to trigger secondary cases at a rate proportional to a predictor η_j of event marks $\boldsymbol{m}_j$. The typical decay of infection pressure with increasing spatial and temporal distance from the infective event is modelled by parametric interaction functions f and g, respectively (Lawson and Leimich, 2000, Section 4). Depending on the application at hand, these could simply be assumed constant, or, e.g., a power-law distance decay for f could be chosen to reflect human travel behaviour on larger scales (Meyer and Held, 2014). The (possibly infinite) upper bounds τ_j and δ_j provide a way of modelling event-specific interaction ranges, but since these have to be specified a common assumption is $\tau_j \equiv \tau$ and $\delta_j \equiv \delta$.

The spatio-temporal point process model (7) corresponds to a branching process with immigration, where part of the event rate is due to the endemic (immigration) component reflecting sporadic cases caused by unobserved sources of infection. The expected number of offspring an event generates according to the "triggering function" $\eta_j\, f(\|\boldsymbol{s}-\boldsymbol{s}_j\|)\, g(t-t_j)$ can thus be interpreted as a model-based effective reproduction number. This number is obtained as the integral of the triggering function over the observed interaction period $(t_j, t_j+\tau_j] \cap (0, T]$ and region $b(\boldsymbol{s}_j, \delta_j) \cap \boldsymbol{W}$, where $b(\boldsymbol{s}_j, \delta_j)$ denotes the disc centered at $\boldsymbol{s}_j$ with radius δ_j.

3.2. Testing for epidemic behaviour

A standard option to assess if the process at hand shows epidemic behaviour is a likelihood ratio test between the full model and the corresponding endemic-only model. The associated test statistic is $D = -2\log(L_{\text{endemic}}/L_{\text{full}})$, where L_{endemic} is the maximized likelihood of the endemic-only model given the observed point pattern, and L_{full} the corresponding likelihood of the full model with an epidemic component. Originally, Meyer et al. (2012) used a log-link for the epidemic predictor, i.e., $\eta_j = \exp(\gamma_0 + \boldsymbol{\gamma}^\top \boldsymbol{m}_j)$, in which case the null hypothesis of an endemic-only model with $\eta_j = 0$ is on the boundary of the parameter space (Self and Liang, 1987). Using the identity link for η_j avoids this problem and allows for the model to represent an inhibition process with a negative epidemic intensity (as long as $\lambda(\boldsymbol{s}, t) \ge 0$ at all $(\boldsymbol{s}, t)$ and $\lambda(\boldsymbol{s}_i, t_i) > 0$ at all events $i = 1, \ldots, n$).

However, the likelihood ratio test is still non-standard since parameters of the interaction functions f and g are not identifiable under the null hypothesis $(\gamma_0, \boldsymbol{\gamma}) = \boldsymbol{0}$. More importantly, we will see that including an epidemic component can improve the fit for reasons other than pure presence of space–time interaction – if the data provide evidence against the piecewise constant endemic intensity. For these reasons, we will determine the null distribution of the test statistic by a Monte Carlo permutation approach just as for the classical significance tests described in Section 2. Note that this is only valid for models with a separable endemic intensity, i.e., independent spatial and temporal background processes. For each of B random relabelings of the event times, the full model and the endemic-only model have to be re-estimated to determine the test statistic for the permuted point pattern. We perform the test conditionally on the estimated interaction functions f and g to avoid identifiability issues for permuted data with a naturally low epidemic intensity, and to drastically reduce the computational burden (the point process likelihood requires the integration of $f(\|\boldsymbol{s}-\boldsymbol{s}_j\|)$ over $\boldsymbol{W} \cap b(\boldsymbol{s}_j, \delta_j)$, see Meyer et al., 2016, Section 3.1).

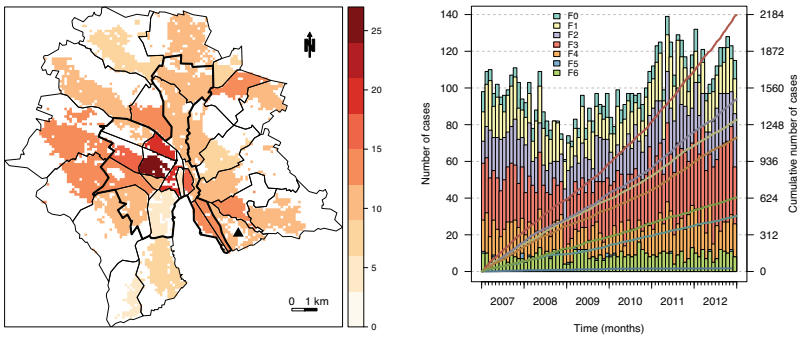
An advantage of the permutational approach is that we can replace the likelihood ratio statistic D with a more accessible quantity, for example the reproduction number described above. Since the expected number of offspring implied by the point process model (7) is event-specific, we use

$$T_R = \hat{\gamma}_0 \left[\int_{b(\boldsymbol{0},\delta)} \hat{f}(\boldsymbol{s})\, \mathrm{d}\boldsymbol{s} \right] \left[\int_0^\tau \hat{g}(t)\, \mathrm{d}t \right] \quad (8)$$

as a basic effect size, i.e., the estimated number of secondary cases triggered by an event with $\boldsymbol{m}_j = \boldsymbol{0}$ during an infectious period of length τ within a surrounding region of radius δ.

4. Psychiatric hospital admissions in Zurich, 2007–2012

We now investigate contagion of psychiatric hospital admissions in an urban catchment area using the tests for space–time interaction described in Section 2 and 3. The spread of the news about a psychiatric hospital admission is thereby assumed to be most likely in close neighbourhood of the patient's residence within a short period of time. Specifically, we assume a maximum interaction radius of $\delta = 250$ metres and an infectious period of $\tau = 14$ days starting from admission. For the main hypothesis, we make no distinction with respect to the patients' diagnoses. However, since contagion might be restricted to a specific subset, we also apply the tests to admissions of



(a) Administrative quarter boundaries of Zurich and their populated area shaded according to the number of patients per 1 000 inhabitants. The black triangle marks the location of the PUK and the thick boundary line encloses inner urban quarters (see text).

(b) Monthly and cumulative number of admissions stratified by diagnosis group.

Fig. 1. Spatio-temporal characteristics of the 7202 admissions to the PUK, 2007–2012.

the two main diagnostic subgroups separately: schizophrenia (ICD-10 F2x) and affective disorders (F3x).

The following section describes the spatio-temporal point pattern of hospital admissions and the additional socio-economic data to be adjusted for in the point process model. Subsequently, the classical and model-based tests for space–time interaction are applied, followed by a discussion of the results.

4.1. Materials

4.1.1. Case reports

We use data of inpatient episodes from the central register of the university hospital of psychiatry (PUK) in Zurich, Switzerland, from the years 2007–2012. The catchment area of the PUK includes about 450,000 to 500,000 inhabitants. The PUK is one of six psychiatric institutions which serve a defined catchment area in the canton of Zurich and which treat the whole spectrum of mental health problems. The PUK covers almost 40% of the treatment episodes of these institutions.

In total the PUK has registered 23,290 admissions of 12,280 different patients during the years 2007–2012. The patient's entry date defines the event time t_i, and the geographic coordinates of the patient's residence in metre units define the corresponding event location $\boldsymbol{s}_i$. We include all voluntary admissions of adult patients living in the city of Zurich at the time of admission, where the primary diagnosis (as qualified at discharge) is in the range F0x–F6x. Of these 10,320 records, we exclude all admissions without proper coordinate information (e.g., unknown/ambiguous address, homeless patients), and all readmissions within 14 days (to not infer self-infections). The remaining sample to analyse consists of 7202 admissions of 4498 patients. Fig. 1 shows the quarter-wise number of patients per 1000 inhabitants as well as the monthly number of admissions.[2] The F2 and F3 subsets contain 1465 (856) and 2180 (1602) admissions (patients), respectively.

4.1.2. Population

The Swiss Federal Statistical Office provided us with population data of the city of Zurich on a 100 m × 100 m (i.e., 1 ha) grid with a total of 392,435 inhabitants. With these data, the observation region $\boldsymbol{W}$ of the point process can be defined as the populated area of Zurich (the shaded area in Fig. 1a). Using the population-based observation region is more adequate than administrative boundaries, which also encompass the large unpopulated areas of the surrounding hills where no events can occur.

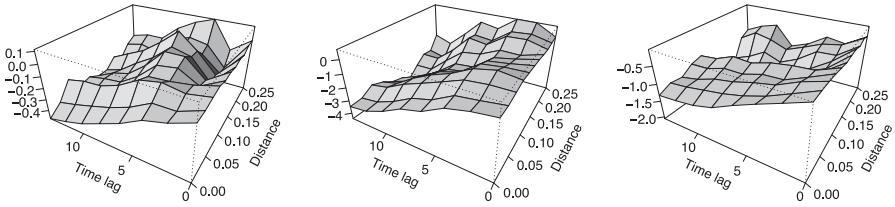
4.1.3. Quarter-specific socio-economic data

Socio-economic characteristics of Zurich's 34 quarters were provided by the Statistics Office of the city of Zurich. Some of the variables were available on a yearly basis but there was not much variation across the years. We thus consistently use the data of the year 2010, which is also the year where most of the time-constant data has been collected as part of that year's census. The following socio-economic characteristics are used: proportion of foreigners, proportion aged 40–64 years, proportion of one-person households, employment rate (proportion employed among the population aged 15 years and over), proportion of inhabitants with minimum (obligatory) education, and mean yearly income per taxpayer. Summary statistics for these quarter-level variables are given in Table 1. An extra indicator variable was constructed from land use statistics to distinguish between urban and rural quarters: each of the inner urban quarters, i.e., those surrounded by a thick line in Fig. 1a, has more than 40% of

[2] Ethical approval for this study was obtained from the ethics committee of the Canton of Zurich, provided that the data will only be published in adequately aggregated form.

Table 1
Summary statistics for socio-economic variables of Zurich's 34 quarters.

	Min.	1st qu.	Median	Mean	3rd qu.	Max.
% Foreigners	20.1	26.0	29.4	29.9	34.4	41.0
% 40–64 years	26.0	29.4	31.4	31.8	33.8	41.4
% One-person households	12.1	20.2	23.4	23.8	26.4	39.1
Employment rate	54.6	62.4	66.1	66.8	71.2	83.2
% Low-level education	14.0	19.9	31.8	29.2	38.3	47.0
Mean income (in 1000 CHF)	35.0	40.8	46.2	47.1	54.0	69.5



(a) All cases ($p = 0.94$). (b) F2 cases ($p = 0.98$). (c) F3 cases ($p = 0.99$).

Fig. 2. Perspective plots of $\hat{D}(\delta, \tau)$ with the permutation-based p-value given in the caption.

its area occupied by buildings and infrastructure and less than 2% forest area.

4.1.4. Density of psychiatrists by city district

The number of psychiatric practices by city district (decomposing Zurich into 12 larger subregions) was obtained from the medical society of the canton of Zurich. To make use of this information when modelling the endemic risk at the smaller quarter level, the number was transformed to a density per hectare of populated district area. This density is then assumed to hold for all quarters of the corresponding district. It ranges from 0.003 to 0.437 practices per hectare. Aggregated over the whole populated area of Zurich, there are 0.08 psychiatric practices per hectare.

4.2. Classical significance tests

Table 2 shows the contingency tables underlying the conducted Knox tests. The associated p-values are 0.79 (all cases), 0.93 (F2), and 0.81 (F3), respectively, suggesting no evidence for space–time interaction at the predefined scales, neither in the overall sample nor within the main diagnoses subgroups.

The standardized Mantel tests yield p-values (Pearson correlations given in parentheses) of 0.092 ($r_{\text{all}} = 0.0044$), 0.13 ($r_{\text{F2}} = 0.0079$), and 0.88 ($r_{\text{F3}} = -0.0069$), respectively. The correlations between spatial and temporal distances are generally very small and even negative for the subgroup of affective disorders. Hence, the Mantel test also yields no evidence for space–time interaction.

Fig. 2 shows perspective plots of $\hat{D}(\delta, \tau)$ from Eq. (5), evaluated on an equidistant grid covering the distance range 0–250 m and the time lags 0–14 days, respectively. For an infectious process one would expect an excess number of "close" cases compared to an interaction-free process. However, this excess as measured by $\hat{D}(\delta, \tau)$ is negative for most interaction ranges and generally decreases with the time lag τ. The omnibus test (6) based on 999 random permutations of the event times yields p-values of 0.94 (all cases), 0.98 (F2 cases), and 0.99 (F3 cases), respectively, all indicating absence of evidence for space–time interaction.

Table 2
Contingency tables of the spatial and temporal distances of all pairs of PUK admissions, and in the subgroups of schizophrenic and affective disorders, respectively. Each table's caption gives the expected number E of close pairs in the absence of space–time interaction and the permutation-based p-value.

(a) All cases ($E = 2448$, $p = 0.79$).

	km apart		
Days apart	≤0.25	>0.25	Σ
≤14	2408	346,953	349,361
>14	179,271	25,402,169	25,581,440
Σ	181,679	25,749,122	25,930,801

(b) F2-cases ($E = 134$, $p = 0.93$).

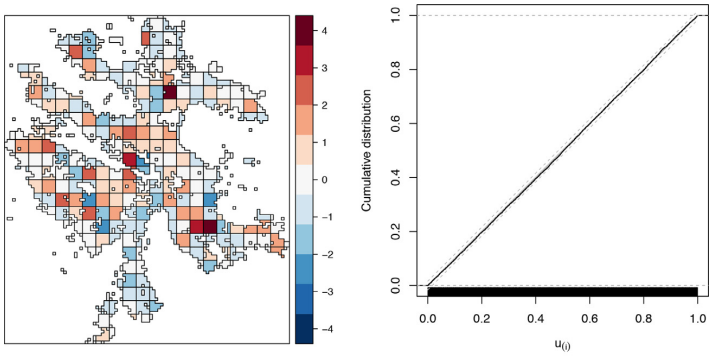
	km apart		
Days apart	≤0.25	>0.25	Σ
≤14	119	14,388	14,507
>14	9774	1,048,099	1,057,873
Σ	9893	1,062,487	1,072,380

(c) F3-cases ($E = 226$, $p = 0.81$).

	km apart		
Days apart	≤0.25	>0.25	Σ
≤14	214	32,564	32,778
>14	16,156	2,326,176	2,342,332
Σ	16,370	2,358,740	2,375,110

4.3. Endemic–epidemic point process model

To simultaneously estimate effects of local socio-economic characteristics on the admission rate and



(a) Pearson residuals on a grid of 500m×500m pixels. Positive values indicate an excess number of admissions compared to what is expected under the fitted model.

(b) Cumulative distribution of the transformed temporal residual process (as in Ogata, 1988, Figure 10) under random sampling of the event times within the corresponding days. The dashed lines show the 95% error bounds of the Kolmogorov-Smirnov statistic.

Fig. 3. Spatial and temporal residual analysis of the fitted model (9), integrating over the other dimension.

investigate additional space–time interaction of the cases, we formulate the following endemic–epidemic point process model:

$$\lambda(\boldsymbol{s},t) = \rho_{[\boldsymbol{s}]} \exp\left(\boldsymbol{\beta}_{\boldsymbol{z}}^{\top}\boldsymbol{z}_{[\boldsymbol{s}]} + \beta_o o_{[\boldsymbol{s}]} + \beta_w \mathbb{1}_w(t)\right) + \gamma_0 \cdot |I(\boldsymbol{s},t)|. \quad (9)$$

The endemic model component accounts for the spatially varying population density $\rho_{[\boldsymbol{s}]}$ and socio-economic characteristics $\boldsymbol{z}_{[\boldsymbol{s}]}$. Additionally, we account for the quarter's distance $o_{[\boldsymbol{s}]}$ to the PUK's quarter (1 for directly adjacent quarters, 2 for second-order neighbours, and so forth), and include an indicator function $\mathbb{1}_w(t)$ for weekends and holidays on which fewer admissions are registered (no scheduled admissions). By the epidemic component, each case is assumed to cause a local increase of γ_0 in the event rate during 14 days in the neighbourhood of 250 m around its residence.

Fitting this model to the point pattern of all 7202 admissions takes only 1.3 s on an ordinary laptop running at 2.80 GHz. The goodness of fit can be evaluated using residual analysis methods for space–time point processes (Clements et al., 2011). Spatial pixel residuals (integrated over time) can help to identify regions with considerably more or less cases than explained by the model (Fig. 3a). The distribution of the transformed residual process in Fig. 3b suggests that the estimated intensity (integrated over space) provides a good description of the temporal point pattern.

Table 3 presents the rate ratio estimates for the endemic effects of model (9). The event rate increases with the proportion of foreigners, inhabitants aged 40–64 years, one-person households, adults employed, and people with minimum education level. It tends to be lower in quarters with a higher mean income and is reduced by 62% during weekends and holidays. There is no evidence for further geographic effects as reflected by the urban indicator, the distance to the PUK and the density of psychiatric practices.

Table 3
Estimated rate ratios (RR) of endemic effects.

	RR	95% CI	*p*-value
Urban quarter	0.991	0.91–1.08	0.84
Distance to PUK	0.983	0.96–1.01	0.16
% Foreigners	1.018	1.01–1.02	<0.0001
% 40–64 years	1.084	1.07–1.10	<0.0001
% One-person households	1.032	1.02–1.04	<0.0001
Employment rate	1.014	1.01–1.02	0.0011
% Low-level education	1.025	1.02–1.03	<0.0001
Mean income (in 1000 CHF)	0.991	0.98–1.00	0.021
Psychiatric practices/ha	1.071	0.67–1.71	0.78
Weekend/holiday	0.379	0.35–0.41	<0.0001

The triggering rate is estimated to be $\hat{\gamma}_0 = 0.023$, which corresponds to $T_R = 0.063$ secondary cases within 14 days and 250 m. The likelihood ratio statistic for $H_0 : \gamma_0 = 0$, i.e., comparing the endemic-only to the full model, is $D = 115.0$. Although the goodness of fit improves considerably with the epidemic component, the permutation test reveals that this improvement does not reflect true space–time interaction. Fig. 4 shows the null distribution of T_R from 199 data sets with permuted event times and thus no space–time interaction by construction. A large proportion $p = 0.87$ of the permuted data sets has a value of T_R *higher* than estimated for the real data. This is in agreement with the results from the classical significance tests in that there is no evidence for epidemicity of psychiatric hospital admissions.

Running the permutation test on the F2 and F3 subsets yields *p*-values of 0.97 and 0.90, respectively. The estimated rate ratios of the endemic effects are displayed in Table 4 and are generally similar in direction and order of magnitude as in the overall model (Table 3). There is a stronger negative association of the event rate with the distance to the PUK, i.e., quarters further apart tend to have a lower rate of admissions (given

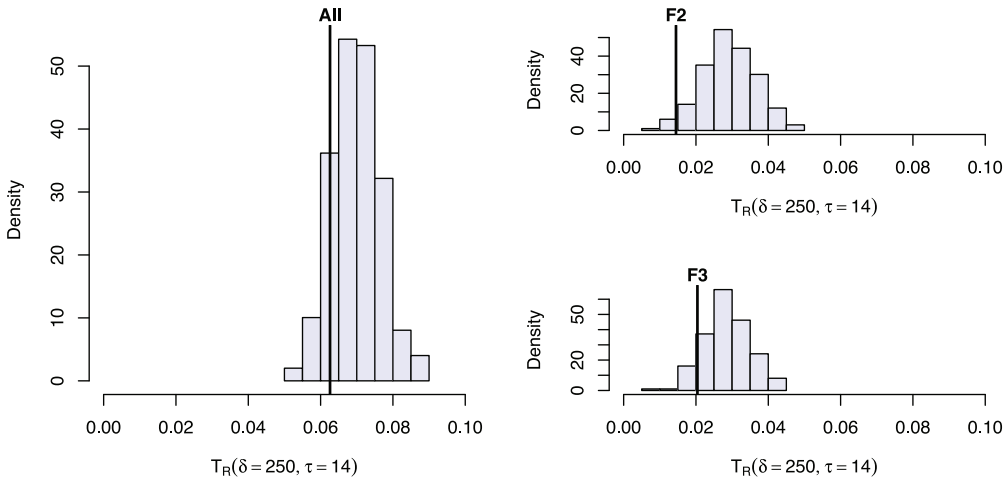


Fig. 4. Permutation distribution of the model-based reproduction number T_R (Eq. (8)).

Table 4
Estimated rate ratios (RR) of endemic effects in the subgroup models.

	F2 cases			F3 cases		
	RR	95% CI	*p*-value	RR	95% CI	*p*-value
Urban quarter	0.852	0.71–1.03	0.094	1.188	1.02–1.38	0.023
Distance to PUK	0.947	0.90–1.00	0.035	0.927	0.89–0.97	0.0003
% Foreigners	1.010	1.00–1.03	0.17	1.003	0.99–1.01	0.61
% 40–64 years	1.140	1.10–1.18	<0.0001	1.050	1.02–1.08	0.0012
% One-person households	1.049	1.03–1.07	<0.0001	1.026	1.01–1.04	0.0003
Employment rate	1.021	1.00–1.04	0.029	1.006	0.99–1.02	0.40
% Low-level education	1.035	1.02–1.05	<0.0001	1.024	1.01–1.04	0.001
Mean income (in 1000 CHF)	0.970	0.95–0.99	0.0002	0.996	0.98–1.01	0.59
Psychiatric practices/ha	1.045	0.38–2.88	0.93	1.052	0.48–2.32	0.90
Weekend/holiday	0.505	0.44–0.58	<0.0001	0.354	0.31–0.40	<0.0001

identical socio-economic characteristics). For the subset of patients with affective disorders, there is no evidence of effects of mean income, proportion of foreigners, and employment rate on the quarter's admission rate, which is, however, estimated to be 19% higher in inner urban quarters than in outer quarters with otherwise similar characteristics.

4.4. Discussion of the results

This observational study shows several quarter-level socio-economic characteristics to be associated with the hospital admission rate. Similar to previous studies suggesting that psychiatric disorders are more frequent in socially deprived areas (Chaix et al., 2006; Reijneveld and Schene, 1998; Sundquist and Ahlen, 2006), lower education and lower income were associated with an increased hospital admission rate. However, the finding that a higher employment rate is (weakly) associated with an increased admission rate is contradictory to previous results (Kammerling and O'Connor, 1993; Simone et al., 2013) and difficult to interpret. Note that we have used the proportion of employed persons among the population aged 15 years and over, whereas other studies used the unemployment rate, which is usually defined among the labour force. For the subset of patients with affective disorders there was no evidence for an effect of the local employment rate. A higher proportion of foreign residents was found to increase the rate of psychiatric admissions, which is in agreement with the recent study by Simone et al. (2013). A higher proportion of one-person households was also found to increase the admission rate, which probably relates to a lack of social support and/or partnership known as a protective factors. However, previous findings on the effect of household size are contradictory (Simone et al., 2013; Torrey and Yolken, 1998).

To test for space–time interaction of psychiatric admissions, we assumed the "spread" to be confined to 250 m with homogeneous infectivity in this neighbourhood. This corresponds to the closest and most influential distance class in a previous study on the spatial clustering of autism (Liu et al., 2010). Regardless of whether a classical test or the model-based test was applied, we found no indication of spatio-temporal interaction of psychiatric hospital admissions in Zurich. We have additionally assessed the sensitivity of the Knox and model-based tests with respect to the assumed upper bounds of spatial and temporal interaction. Restricting interaction to cases in the same

building only ($\delta = 5$ m) or to $\delta = 50$ or 500 m and/or assuming shorter infectious periods of $\tau = 7$ or 3 instead of 14 days led to similar test results with no evidence for space–time interaction. For instance, the smallest p-values obtained for the F2 subset are 0.28 (Knox) and 0.22 (model-based), respectively, for the thresholds $\delta =$ 250 m and $\tau = 3$ days corresponding to the highest value of $\hat{D}(\delta, \tau)$ in Fig. 2b

The spatial interaction function can be seen as a rough proxy for the population's social contact network, which is the natural driver of person-to-person transmission. However, the potential interaction between subsequent admissions through social contacts might not map well into spatially confined clusters as suggested by the point process model. One potentially missing link is through the place of work, where colleagues might get informed about the hospital stay. Although we find no evidence of clustering of psychiatric admissions on the spatio-temporal scale, interaction between cases might become apparent if the social contact structure could be taken into account (Christakis and Fowler, 2013; Coviello et al., 2014).

Another limitation of this analysis of inpatient episodes from a single hospital is underreporting, since "infected" neighbours might not be admitted to the PUK in the first place. They might turn to a general practitioner, a registered psychiatrist, or another psychiatric institution, and may or may not be eventually admitted to the PUK. Similar to missing links via long-range social contacts, such unobserved cases lead to missing clusters and underestimation of spatio-temporal interaction via the epidemic component.

To quantify the strength of interaction, the point process model enables the estimation of an effective reproduction number, which is also used as test statistic T_R. However, the value of T_R has to be interpreted with caution since it is affected by the endemic formulation. Specifically, the permutation test revealed that a positive force of infection is estimated even under absence of space–time interaction. This happens because the data do not support the piecewise constant endemic intensity on the chosen grid. Although we have restricted the observation region to the populated area of Zurich, the population density actually varies within each quarter. Given the large amount of 7202 cases, there is evidence that subsequent cases tend to be closer to a previous case than implied by a spatially homogeneous intensity simply due to spatial clustering of the population. The permutation test accounts for this endemic misspecification in the null distribution of the test statistic by keeping the marginal spatial and temporal locations fixed. If significant space–time interaction is found (as for the occurrence of invasive meningococcal disease, see Supplementary material), the difference between the observed T_R and its average value under permutation could be used as an effect size quantifying true space–time interaction. Note that for data conforming to the piecewise constant endemic intensity, for example, for simulations from the fitted endemic–epidemic point process model, the null distribution of T_R is centred around zero (see Supplementary material).

5. Conclusions

Classical significance tests of space–time clustering solely operate on the events' spatial and temporal distance matrices to evaluate the association of spatial and temporal closeness. In this paper we have demonstrated how spatio-temporal interaction can be investigated more thoroughly using a two-component point process regression model. While its endemic component explains the background risk of new events through regional and/or time-varying covariates, the superposed epidemic component reflects that events tend to appear in spatio-temporal clusters. Hence, the model-based global clustering test is adjusted for, e.g., socio-economic heterogeneity and seasonal effects.

For the psychiatric hospital admissions in Zurich, we found several quarter-level socio-economic characteristics to be associated with the endemic occurrence of cases. However, there was no evidence for spatio-temporal interaction of the cases. The hypothesized social influence on help-seeking behaviour is thus not apparent in local clusters of sequential hospital admissions. Therefore, standard methods for non-contagious spatial processes could be used to analyse these data, e.g., spatial point pattern analysis, or disease mapping and ecological regression approaches (Baddeley et al., 2015; Waller and Gotway, 2004). A simple example is a Poisson regression model for the number of patients by quarter using the population as a multiplicative offset and the socio-economic characteristics as explanatory variables (see, e.g., Lawson, 2013, Section 7.7, Marshall, 1991, or Wakefield, 2007). In mathematical terms, the expected number of patients (the Poisson rate) of quarter i could be modelled as $\lambda_i = n_i \cdot \exp(\boldsymbol{\beta}^\top \boldsymbol{z}_i)$, where n_i is the quarter's population and $\boldsymbol{z}_i$ the vector of socio-economic characteristics. Doing so, the estimated risk ratios and confidence intervals are in fact very similar to those from the endemic–epidemic point process model presented in Table 3. Alternatively, a separable *endemic-only* point process model could also serve as input for further exploratory analysis of second-order properties based on the inhomogeneous K-function of Gabriel and Diggle (2009), which requires an estimate of the intensity $\lambda(\boldsymbol{s}_i, t_i)$ at all events.

If no individual cases but only area-level counts over several time periods are available, the tests for spatial, temporal and spatio-temporal clustering proposed by Raubertas (1988) could be used. The areal time-series model of Meyer and Held (2014, Section 3) also offers means of quantifying the importance of transmission from neighbouring regions and estimating the spatial distance decay of interaction.

The methods discussed in this paper are "general tests of clustering" not to be confused with "tests for the detection of clusters" (Besag and Newell, 1991). Kulldorff (2006) reviews both of these for purely spatial processes. Methods to detect emergent space–time clusters have been proposed by, e.g., Piroutek et al. (2014) monitoring the cumulative sum of a local Knox statistic, or Kulldorff et al. (2005) using a space–time scan statistic.

Although the presented endemic–epidemic point process regression framework is not intended to replace the convenient classical tests, its added value is manifold: first, it offers insight into endemic characteristics of the process via modelling the background rate using covariates for which the test is adjusted. Second, a distance decay of interaction can be estimated and incorporated into the test. An example is the power-law kernel that we have used to reflect large-scale human travel in modelling the occurrence of invasive meningococcal disease in Germany (Meyer and Held, 2014). For comparison, we have also applied the various test procedures to these data and find clear evidence of spatio-temporal interaction as expected for an infectious disease (see Supplementary material). Last but not least, if there is evidence for space–time interaction, the model allows for a more detailed estimation of event-specific infectiousness through the epidemic predictor η_j. The local population density or event characteristics, such as the patient's age group or disease severity, could be associated with the triggering rate, such that events are expected to produce a varying number of secondary cases. Whether the model-based test also has more power to detect space–time interaction than the classical tests – especially if "mechanistic" knowledge about the process and suitable covariates enable a well-fitting model – is an open question to be possibly answered by future simulation studies.

Acknowledgements

We thank Michael Höhle (Stockholm University) and two anonymous referees for illuminating comments on a previous version of this manuscript, as well as Dominik Ullmann (Swiss Federal Statistical Office) and Michael Böniger (City of Zurich Statistics Office) for providing the socio-economic data of Zurich's quarters. The work of the first author was financially supported by the Swiss National Science Foundation (Project #137919).

Appendix A. Software

All analyses were performed using the statistical software environment R (R Core Team, 2015). We have implemented the model-based test (`epitest`) as well as the Knox test (`knox`) in the framework of the R package surveillance (Höhle et al., 2016). The package contains methods for visualization, inference and simulation of the endemic–epidemic point process model (as described in Meyer et al., 2016, Section 3), as well as a wrapper (`stKtest`) to perform *K*-function analysis using the splancs package (Rowlingson and Diggle, 2015). For all these tests, our implementations allow the *B* permutations to be distributed across multiple cores. To perform the standardized Mantel test, we used a C implementation (`mantel.randtest`) from the package ade4 (Chessel et al., 2004) as at version 1.7-4.

Embedding the test for interaction in a point process regression model for the conditional intensity function comes at the cost of a considerably increased runtime compared to the classical tests (Table A.5). Fitting a single model (with constant spatial interaction function) only takes about a second for a given data set, but the permutation approach requires the two competing models (with and without the epidemic component) to be re-estimated on each of the *B* permutations. Furthermore, the extra-long runtime for the whole data set (also for the Knox test) is due to memory-intensive distance matrix calculations, which do not scale well on a laptop. This will, however, very much improve by a more efficient implementation of the permutation handling planned for future versions of the R package surveillance.

Table A.5
Runtime comparison of the various tests for space–time interaction. The timings refer to the real elapsed time in minutes. Note that a C implementation was used for the Mantel test; for the other tests, the *B* permutations were distributed across four cores ($B = 999$ for the classical tests and $B = 199$ for the model-based approach).

	Knox	Mantel	*K*-function	Model-based
All cases ($n = 7202$)	20.6	1.7	8.8	44.0
F2 cases ($n = 1465$)	0.6	0.1	0.3	1.6
F3 cases ($n = 2180$)	1.6	0.2	0.7	3.0

Shapefiles have been edited using QGIS 2.2 (Quantum GIS Development Team, 2014) with the fTools plugin 0.6.2. The original administrative boundaries have been simplified using MapShaper.org v. 0.1.18 (Harrower and Bloch, 2006) to speed up computations. Within R, the packages sp (Pebesma and Bivand, 2005), rgdal (Bivand et al., 2015), rgeos (Bivand and Rundel, 2016), and spatstat (Baddeley et al., 2015) were used to deal with geographic shapes. The list of holidays in Zurich was obtained from package timeDate (Wuertz, 2015). Tables have been created with the xtable package (Dahl, 2016). This manuscript has been generated dynamically in R version 3.2.3 (2015-12-10) using knitr (Xie, 2015).

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.sste.2016.03.002.

References

Baddeley A, Rubak E, Turner R. Chapman & Hall/CRC interdisciplinary statistics. Spatial point patterns: methodology and applications with R. Chapman and Hall/CRC; 2015.
Balderama E, Schoenberg FP, Murray E, Rundel PW. Application of branching models in the study of invasive species. J Am Stat Assoc 2012;107(498):467–76. doi:10.1080/01621459.2011.641402.
Besag J, Diggle PJ. Simple Monte Carlo tests for spatial pattern. J R Stat Soc Ser C: Appl Stat 1977;26(3):327–33. doi:10.2307/2346974.
Besag J, Newell J. The detection of clusters in rare diseases. J R Stat Soc Ser A: Stat Soc 1991;154:143–55. doi:10.2307/2982708.
Birch JM, Alexander FE, Blair V, Eden OB, Taylor GM, McNally RJQ. Space–time clustering patterns in childhood leukaemia support a role for infection. Br J Cancer 2000;82(9):1571–6. doi:10.1054/bjoc.1999.1072.
Bivand R., Keitt T., Rowlingson B.. rgdal: bindings for the geospatial data abstraction library. R package version 1.1-3; 2015.
Bivand R., Rundel C.. rgeos: interface to geometry engine – open source (GEOS). R package version 0.3-17; 2016.
Chaix B, Leyland AH, Sabel CE, Chauvin P, Råstam L, Kristersson H, Merlo J. Spatial clustering of mental disorders and associated characteristics of the neighbourhood context in Malmö, Sweden, in 2001. J Epidemiol Commun Health 2006;60(5):427–35. doi:10.1136/jech.2005.040360.
Chessel D, Dufour AB, Thioulouse J. The ade4 package – I: one-table methods. R News 2004;4(1):5–10.

Christakis NA, Fowler JH. Social contagion theory: examining dynamic social networks and human behavior. Stat Med 2013;32(4):556–77. doi:10.1002/sim.5408.
Clements RA, Schoenberg FP, Schorlemmer D. Residual analysis methods for space–time point processes with applications to earthquake forecast models in California. Ann Appl Stat 2011;5(4):2549–71. doi:10.1214/11-AOAS487.
Coviello L, Sohn Y, Kramer ADI, Marlow C, Franceschetti M, Christakis NA, Fowler JH. Detecting emotional contagion in massive social networks. PLoS One 2014;9(3):e90315. doi:10.1371/journal.pone.0090315.
Dahl D.B. xtable: export tables to LaTeX or HTML. R package version 1.8-2; 2016.
Daley DJ, Gani J. Epidemic modelling: an introduction. Cambridge University Press; 1999.
Diggle PJ. Monographs on statistics & applied probability. Statistical analysis of spatial and spatio-temporal point patterns. 3rd ed. Boca Raton: Chapman & Hall/CRC; 2013.
Diggle PJ, Chetwynd AG, Häggkvist R, Morris SE. Second-order analysis of space–time clustering. Stat Methods Med Res 1995;4(2):124–36. doi:10.1177/096228029500400203.
Diggle PJ, Rowlingson B, Su Tl. Point process methodology for on-line spatio-temporal disease surveillance. Environmetrics 2005;16(5):423–34. doi:10.1002/env.712.
Eckley DC, Curtin KM. Evaluating the spatiotemporal clustering of traffic incidents. Comput Environ Urban Syst 2013;37:70–81. doi:10.1016/j.compenvurbsys.2012.06.004.
Fowler JH, Christakis NA. Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the Framingham Heart Study. BMJ 2008;337:a2338. doi:10.1136/bmj.a2338.
Gabriel E, Diggle PJ. Second-order analysis of inhomogeneous spatio-temporal point process data. Stat Neerl 2009;63(1):43–51. doi:10.1111/j.1467-9574.2008.00407.x.
Grubesic TH, Mack EA. Spatio-temporal interaction of urban crime. J Quant Criminol 2008;24(3):285–306. doi:10.1007/s10940-008-9047-5.
Gustafsson B, Carstensen J. Space–time clustering of childhood lymphatic leukaemias and non-Hodgkin's lymphomas in Sweden. Eur J Epidemiol 2000;16(12):1111–16. doi:10.1023/A:1010953713048.
Harrower M, Bloch M. Mapshaper.org: a map generalization web service. IEEE Comput Graph Appl 2006;26(4):22–7. doi:10.1109/MCG.2006.85.
Haw C, Hawton K, Niedzwiedz C, Platt S. Suicide clusters: a review of risk factors and mechanisms. Suicide Life Threat Behav 2013;43(1):97–108. doi:10.1111/j.1943-278X.2012.00130.x.
Held L, Höhle M, Hofmann M. A statistical framework for the analysis of multivariate infectious disease surveillance counts. Stat Model 2005;5(3):187–99. doi:10.1191/1471082X05st098oa.
Höhle M., Meyer S., Paul M.. surveillance: temporal and spatio-temporal modeling and monitoring of epidemic phenomena. R package version 1.11.0; 2016.
Jacquez GM. A *k* nearest neighbour test for space–time interaction. Stat Med 1996;15(18):1935–49. doi:10.1002/(SICI)1097-0258(19960930)15:18<1935::AID-SIM406>3.0.CO;2-I.
Johnson SD. A brief history of the analysis of crime concentration. Eur J Appl Math 2010;21(4–5):349–70. doi:10.1017/S0956792510000082.
Kammerling RM, O'Connor S. Unemployment rate as predictor of rate of psychiatric admission. BMJ 1993;307:1536–9. doi:10.1136/bmj.307.6918.1536.
Knox EG. The detection of space–time interactions. J R Stat Soc Ser C: Appl Stat 1964;13(1):25–30. doi:10.2307/2985220.
Knox G. Detection of low intensity epidemicity: application to cleft lip and palate. Br J Prev Soc Med 1963;17(3):121–7.
Kulldorff M. Tests of spatial randomness adjusted for an inhomogeneity. J Am Stat Assoc 2006;101(475):1289–305. doi:10.1198/016214506000000618.
Kulldorff M, Heffernan R, Hartman J, AssunÈºÈúo R, Mostashari F. A space–time permutation scan statistic for disease outbreak detection. PLoS Med 2005;2(3):e59. doi:10.1371/journal.pmed.0020059.
Kulldorff M, Hjalmars U. The Knox method and other tests for space–time interaction. Biometrics 1999;55(2):544–52. doi:10.1111/j.0006-341X.1999.00544.x.
Lawson AB. Bayesian disease mapping: hierarchical modeling in spatial epidemiology. 2nd. Chapman & Hall/CRC; 2013.
Lawson AB, Leimich P. Approaches to the space–time modelling of infectious disease behaviour. IMA J Math Appl Med Biol 2000;17(1):1–13. doi:10.1093/imammb/17.1.1.
Liu KY, King M, Bearman PS. Social influence and the autism epidemic. AJS 2010;115(5):1387–434. doi:10.1086/651448.
Mack EA, Malizia N, Rey SJ. Population shift bias in tests of space–time interaction. Comput Environ Urban Syst 2012;36(6):500–12. doi:10.1016/j.compenvurbsys.2012.05.001.
Mantel N. The detection of disease clustering and a generalized regression approach. Cancer Res 1967;27(2):209–20.
Marshall RJ. A review of methods for the statistical analysis of spatial patterns of disease. J R Stat Soc Ser A: Stat Soc 1991;154(3):421–41. doi:10.2307/2983152.
McNally RJ, Alexander FE, Bithell JF. Space–time clustering of childhood cancer in great Britain: a national study, 1969–1993. Int J Cancer 2006;118(11):2840–6. doi:10.1002/ijc.21726.
Meyer S, Elias J, Höhle M. A space–time conditional intensity model for invasive meningococcal disease occurrence. Biometrics 2012;68(2):607–16. doi:10.1111/j.1541-0420.2011.01684.x.
Meyer S, Held L. Power-law models for infectious disease spread. Ann Appl Stat 2014;8(3):1612–39. doi:10.1214/14-AOAS743.
Meyer S, Held L, Höhle M. Spatio-temporal analysis of epidemic phenomena using the R package surveillance. J Stat Softw 2016. [in press]. http://arxiv.org/abs/1411.0416
Mohler GO, Short MB, Brantingham PJ, Schoenberg FP, Tita GE. Self-exciting point process modeling of crime. J Am Stat Assoc 2011;106(493):100–8. doi:10.1198/jasa.2011.ap09546.
Morris SE, Zelner JL, Fauquier DA, Rowles TK, Rosel PE, Gulland F, Grenfell BT. Partially observed epidemics in wildlife hosts: modelling an outbreak of dolphin morbillivirus in the northwestern Atlantic, June 2013–2014. J R Soc Interface 2015;12(112):20150676. doi:10.1098/rsif.2015.0676.
Møller J, Ghorbani M. Aspects of second-order analysis of structured inhomogeneous spatio-temporal point processes. Stat Neerl 2012;66(4):472–91. doi:10.1111/j.1467-9574.2012.00526.x.
Ogata Y. Space–time point-process models for earthquake occurrences. Ann Inst Stat Math 1998;50(2):379–402. doi:10.1023/A:1003403601725.
Pebesma EJ, Bivand RS. Classes and methods for spatial data in R. R News 2005;5(2):9–13.
Piroutek A, Assunção R, Paiva T. Space–time prospective surveillance based on Knox local statistics. Stat Med 2014;33(16):2758–73. doi:10.1002/sim.6118.
Quantum GIS Development Team. Quantum GIS geographic information system. Open source geospatial foundation. QGIS version 2.2.0; 2014.
R Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2015.
Raubertas RF. Spatial and temporal analysis of disease occurrence for detection of clustering. Biometrics 1988;44(4):1121–9. doi:10.2307/2531740.
Reijneveld SA, Schene AH. Higher prevalence of mental disorders in socioeconomically deprived urban areas in The Netherlands: community or personal disadvantage? J Epidemiol Community Health 1998;52(1):2–7. doi:10.1136/jech.52.1.2.
Rosenquist JN, Fowler JH, Christakis NA. Social network determinants of depression. Mol Psychiatry 2011;16(3):273–81. doi:10.1038/mp.2010.13.
Rowlingson B., Diggle P.. splancs: Spatial and space–time point pattern analysis. R package version 2.01-38; 2015.
Self SG, Liang KY. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. J Am Stat Assoc 1987;82(398):605–10. doi:10.1080/01621459.1987.10478472.
Simone C, Carolin L, Max S, Reinhold K. Associations between community characteristics and psychiatric admissions in an urban area. Soc Psychiatry Psychiatr Epidemiol 2013;48(11):1797–808. doi:10.1007/s00127-013-0667-1.
Sundquist K, Ahlen H. Neighbourhood income and mental health: a multilevel follow-up study of psychiatric hospital admissions among 4.5 million women and men. Health Place 2006;12(4):594–602. doi:10.1016/j.healthplace.2005.08.011.
Tango T. Statistics for biology and health. Statistical methods for disease clustering. Springer; 2010.
Torrey EF, Yolken RH. Is household crowding a risk factor for schizophrenia and bipolar disorder? Schizophr Bull 1998;24(3):321–4.
Wakefield J. Disease mapping and spatial regression with count data. Biostatistics 2007;8(2):158–83. doi:10.1093/biostatistics/kxl008.
Waller LA, Gotway CA. Wiley series in probability and statistics. Applied spatial statistics for public health data. John Wiley & Sons; 2004.
Ward MP, Carpenter TE. Analysis of time–space clustering in veterinary epidemiology. Prev Vet Med 2000;43(4):225–37. doi:10.1016/S0167-5877(99)00111-7.
Wuertz D.. timeDate: Rmetrics – chronological and calendar objects. R package version 3012.100; 2015.
Xie Y. The R series. Dynamic documents with R and knitr. 2nd ed. Boca Raton, Florida: Chapman and Hall/CRC; 2015.

Supplement to "Model-based testing for space-time interaction using point processes"

Sebastian Meyer[a,*], Leonhard Held[a]

[a] *Epidemiology, Biostatistics and Prevention Institute (EBPI), University of Zurich, Hirschengraben 84, 8001 Zürich, Switzerland*

Abstract

In this supplement, we apply the various tests for space-time interaction discussed in the main manuscript to invasive meningococcal disease occurrence in Germany, 2002–2008, as analysed by Meyer et al. (2012) in their original proposal of the spatio-temporal endemic-epidemic point process model. Note that we refer to equations of the main paper via double parentheses, e.g., ((1)), and to equations in the present document using single parentheses.

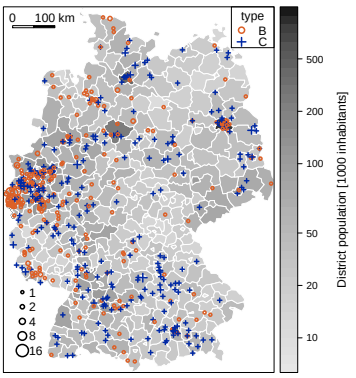
1. Introduction

Invasive meningococcal disease (IMD) is an infectious disease transmitted by large droplet secretions from the respiratory tract of colonized or infected humans (Rosenstein et al., 2001). As such, individual cases are intrinsically dependent and form spatio-temporal clusters, which should be detected by the interaction tests. The temporal range of interaction is determined by the generation time (*aka* serial interval) of the disease, i.e., the time between infection of an individual and one of its secondary cases. As in the original analysis, we use $\tau = 30$ days as an upperbound for the temporal extent of a cluster (Zangwill et al., 1997). The spatial scale of interaction is essentially influenced by the movements of infectious people. Although there are occasional long-range transmissions, we assume most clusters to be confined to $\delta = 50$ kilometres. This threshold is only crucial for the Knox test, all other tests do not strictly depend on such a single critical value.

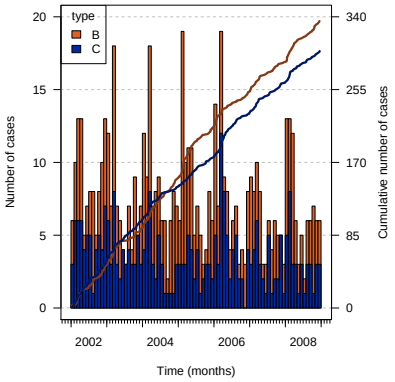
The data consists of 635 cases caused by the two most common meningococcal finetypes in Germany from 2002 to 2008, *B:P1.7-2,4:F1-5* (335 cases) and *C:P1.5,2:F3-3* (300 cases), as reported to and typed by the German Reference Centre for Meningococci. Figure 1 shows the spatial and temporal patterns of the registered cases. Note that cases of different types are not directly related, i.e., the data actually covers epidemics of two distinct pathogens.

*Corresponding author

Email addresses: sebastian.meyer@uzh.ch (Sebastian Meyer), leonhard.held@uzh.ch (Leonhard Held)



(a) Spatial pattern with dot size proportional to the number of cases (postcode level). The outlined western districts form the "Greater Aachen" hotspot of type B (Elias et al., 2010).



(b) Monthly aggregated time series and evolution of the cumulative number of cases (by date of specimen sampling).

Figure 1: Distribution of the 635 IMD cases in Germany, 2002–2008, caused by the two most common meningococcal finetypes B:P1.7-2,4:F1-5 (335 cases) and C:P1.5,2:F3-3 (300 cases).

2. Classical tests for space-time interaction

Table 1: Contingency tables of the spatial and temporal distances of all pairs of IMD cases, stratified by finetype. The expected numbers of close pairs in the absence of space-time interaction are 185 and 49, respectively.

(a) B-type.

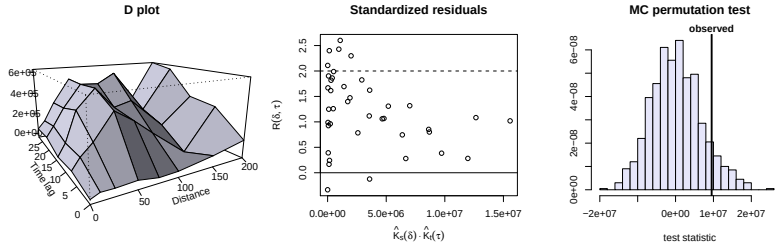
days apart	km apart ≤ 50	km apart > 50	$\sum$
≤ 30	206	1310	1516
> 30	6606	47823	54429
$\sum$	6812	49133	55945

(b) C-type.

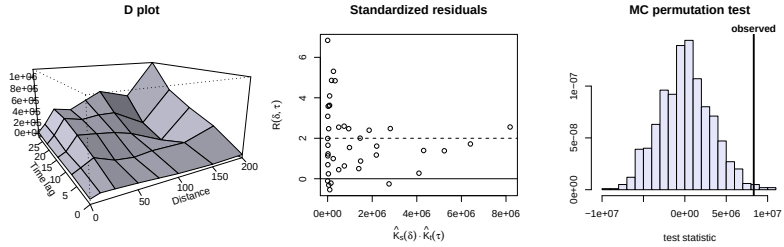
days apart	km apart ≤ 50	km apart > 50	$\sum$
≤ 30	72	1081	1153
> 30	1828	41869	43697
$\sum$	1900	42950	44850

Table 1 shows the contingency tables underlying the Knox test for both meningococcal finetypes. The associated p-values are 0.047 (B) and 0.002 (C), respectively. The standardized Mantel test yields p-values of 0.471 ($r_B = 0.00059$) and 0.001 ($r_C = 0.072$), respectively. Both tests provide strong evidence for spatio-temporal clustering of the C-type cases, while this is less pronounced for the B-type.

For the space-time K-function analysis we vary the temporal distance τ from 0 to 4 weeks and the spatial distance δ over the grid 0, 10, 25, 50, 75, 100, 150 and 200 kilometres. Figure 2 shows the resulting diagnostic plots as proposed by Diggle et al. (1995). As for the other tests, we observe stronger interaction for the C-type, but now the plot of $\hat{D}(\delta, \tau)$ also allows us to qualify the increase in the number of close cases attributable to space-time interaction as a function of the critical distances. Whereas the B-type cases appear to be strongly clustered within roughly 50 kilometres and two weeks, the spatio-temporal clusters of



(a) B-type.



(b) C-type.

Figure 2: Diagnostic plots of the space-time K-function analysis stratified by finetype.

C-type cases seem to have a larger extent. This agrees with the observation of Meyer et al. (2012) that type B cases exhibit a more stationary pattern, especially in the Greater Aachen region (Elias et al., 2010). The omnibus test ((6)) over the selected spatio-temporal scales yields p-values of 0.095 (B) and 0.008 (C), respectively.

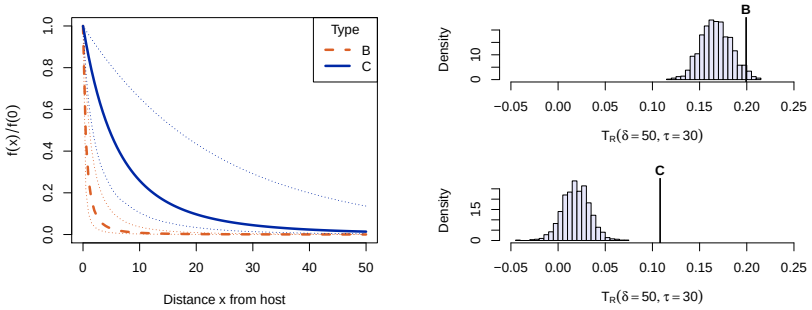
3. Model-based assessment of space-time interaction

With regard to the population dependence and the seasonal pattern visible in Figure 1, we use the endemic model formulation

$$\rho_{[\boldsymbol{s}]} \cdot \exp\left(\beta_0 + \beta_{\text{trend}}[t] + \beta_{\sin}\sin(\omega[t]) + \beta_{\cos}\cos(\omega[t])\right),$$

where $\rho_{[\boldsymbol{s}]}$ is the population density in the district of location $\boldsymbol{s}$ and $\omega = 2\pi/365$ is the frequency of the sinusoidal wave (Held and Paul, 2012). For the force of infection, we apply the unbounded power-law kernel $f(x) = (x + \sigma)^{-d}$ for the spatial distance x, and the simple constant function $g(t) = \mathbb{1}(t \leq 30)$ for the time course of infectivity as in Meyer and Held (2014). In the joint model of Meyer and Held (2014), these interaction functions were assumed to be the same for both finetypes since the power-law decay should reflect human travel behaviour regardless of the specific bacterial agent. Here we estimate separate finetype-specific point process models to test for space-time interaction. Thus the spatial interaction function can adjust for the degree of clustering observed in the different point patterns.

The estimated finetype-specific power laws are displayed in Figure 3a. The stronger distance decay for type B confirms the findings from the space-time



(a) Power-law distance decay of the force of infection (with 95% confidence intervals).

(b) Permutation distribution of T_R ((8)) for types B (top) and C (bottom), respectively.

Figure 3: Estimates from the finetype-specific IMD models.

K-function analysis. While the force of infection is halved already at a distance of only 0.43 km for type B, this value is 4.7 km for type C. The estimated numbers T_R ((8)) of secondary cases within 50 kilometres and 30 days are 0.199 ($p = 0.028$) for type B and 0.108 ($p = 0.001$) for type C (Figure 3b). For both meningococcal finetypes there is evidence for spatio-temporal interaction as expected for an infectious disease. In contrast to the classical approaches, this model-based test is adjusted for seasonality and population heterogeneity via the estimated endemic intensity, and uses an estimated force of infection which is smoothly decreasing with distance.

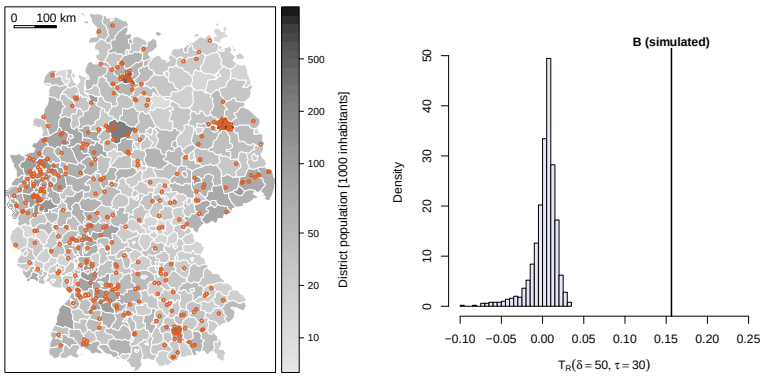
However, the permutation test also reveals a model deficiency especially for type B: the distribution of T_R is centered far above 0 under permutation (Figure 3b), i.e., even for permuted data the fit is improved if an epidemic component is included. This is most likely caused by the simplifying assumption of a locally constant endemic intensity in each district. Since the B-type is persistent in the Greater Aachen region all along the study period, the endemic misspecification with regard to the underlying population becomes important and is corrected towards more clustered event locations by the epidemic component. If we performed the permutation test on data conforming to the actual endemic formulation, for example, on data simulated from the fitted endemic-epidemic point process model, the null distribution would be centered around 0 (Figure 4 illustrates this for a simulation from the fitted model for type B).

The difference between the observed T_R and the mean value under permutation could thus be regarded as a more suitable measure of interaction (0.032 for type B and 0.089 for type C).

References

Diggle PJ, Chetwynd AG, Häggkvist R, Morris SE. Second-order analysis of space-time clustering. Stat Methods Med Res 1995;4(2):124–36. doi:10.1177/096228029500400203.

Elias J, Schouls LM, van de Pol I, Keijzers WC, Martin DR, Glennie A, Oster P, Frosch M, Vogel U, van der Ende A. Vaccine preventability of meningococcal clone, Greater Aachen Region, Germany. Emerg Infect Dis 2010;16(3):465–72. doi:10.3201/eid1603.091102.



(a) Spatial distribution.

(b) Monte Carlo permutation test.

Figure 4: Simulation from the fitted model for type B (359 events).

Held L, Paul M. Modeling seasonality in space-time infectious disease surveillance data. Biom J 2012;54(6):824–43. doi:10.1002/bimj.201200037.

Meyer S, Elias J, Höhle M. A space-time conditional intensity model for invasive meningococcal disease occurrence. Biometrics 2012;68(2):607–16. doi:10.1111/j.1541-0420.2011.01684.x.

Meyer S, Held L. Power-law models for infectious disease spread. Ann Appl Stat 2014;8(3):1612–39. doi:10.1214/14-AOAS743.

Rosenstein NE, Perkins BA, Stephens DS, Popovic T, Hughes JM. Meningococcal disease. N Engl J Med 2001;344(18):1378–88. doi:10.1056/NEJM200105033441807.

Zangwill KM, Schuchat A, Riedo FX, Pinner RW, Koo DT, Reeves MW, Wenger JD. School-based clusters of meningococcal disease in the United States. JAMA 1997;277:389–95. doi:10.1001/jama.1997.03540290041030.

APPENDIX A

A space-time conditional intensity model for invasive meningococcal disease occurrence

Sebastian Meyer, Johannes Elias, Michael Höhle

Published in *Biometrics*, 2012, **68** (2), 607–616.

Biometrics 68, 607–616
June 2012

DOI: 10.1111/j.1541-0420.2011.01684.x

A Space–Time Conditional Intensity Model for Invasive Meningococcal Disease Occurrence

Sebastian Meyer,[1,2,*] Johannes Elias,[3] and Michael Höhle[2,4,]**

[1]Department of Psychiatry and Psychotherapy, Ludwig-Maximilians-Universität, 80336 München, Germany
[2]Department of Statistics, Ludwig-Maximilians-Universität, 80539 München, Germany
[3]German Reference Centre for Meningococci, University of Würzburg, 97080 Würzburg, Germany
[4]Department for Infectious Disease Epidemiology, Robert Koch Institute, 13086 Berlin, Germany
[*]*email:* Sebastian.Meyer@med.uni-muenchen.de
[**]*email:* HoehleM@rki.de

Summary. A novel point process model continuous in space–time is proposed for quantifying the transmission dynamics of the two most common meningococcal antigenic sequence types observed in Germany 2002–2008. Modeling is based on the conditional intensity function (CIF), which is described by a superposition of additive and multiplicative components. As an epidemiological interesting finding, spread behavior was shown to depend on type in addition to age: basic reproduction numbers were 0.25 (95% CI 0.19–0.34) and 0.11 (95% CI 0.07–0.17) for types B:P1.7–2,4:F1–5 and C:P1.5,2:F3–3, respectively. Altogether, the proposed methodology represents a comprehensive and universal regression framework for the modeling, simulation, and inference of self-exciting spatiotemporal point processes based on the CIF. Usability of the modeling in biometric practice is promoted by an implementation in the `R` package `surveillance`.

Key words: Conditional intensity function; Infectious disease surveillance data; Spatiotemporal point process; Stochastic epidemic modeling.

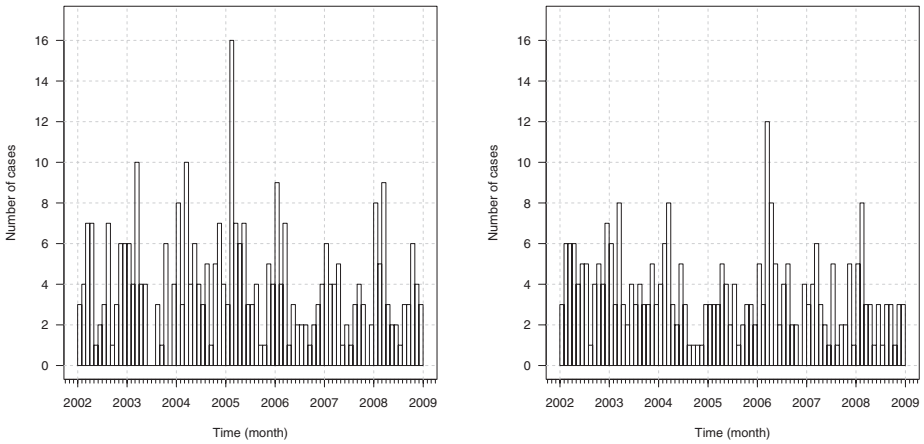
1. Introduction

Infectious diseases—such as influenza, gastroenteritis, and the "swine flu" among humans, or foot and mouth disease, the "bird flu," and classical swine fever among animals—are a matter of tremendous public concern especially gaining attention in case of outbreaks. The present work concentrates on stochastic modeling and associated inference for spatiotemporal epidemic point referenced data motivated by the analysis of routinely collected invasive meningococcal disease (IMD) data. IMD is a life-threatening human bacterial disease mostly manifesting as meningitis or sepsis. Its pathogenic agent, *Neisseria meningitidis* (aka *meningococcus*), can be transmitted by large droplet secretions from the respiratory tract of colonized or infected humans. The only reservoir of meningococci is the human (mostly nasopharyngeal) mucosa (Rosenstein et al., 2001). Data on cases of IMD related to the two most common meningococcal finetypes B:P1.7-2,4:F1-5 and C:P1.5,2:F3-3 in Germany 2002–2008 are obtained from the German Reference Centre for Meningococci (Nationales Referenzzentrum für Meningokokken, NRZM). Here, a "finetype" represents a unique combination of serogroup, sequence type of variable regions 1 and 2 of the outer membrane protein PorA, and sequence type of the variable region of the outer membrane protein FetA. One specific question of interest for the researchers at the NRZM is whether the two finetypes (in what follows abbreviated B and C) exhibit different spatiotemporal behavior.

The postal code of the patient's home address was the spatial resolution available for our analysis. Despite being spatially discrete we consider centroids of postal code areas as quasi-continuous in space when looking at entire Germany. As usual with infectious diseases, the actual time point of infection is unknown for the IMD cases. Therefore, we define the beginning of illness and infectivity as the date of specimen sampling.

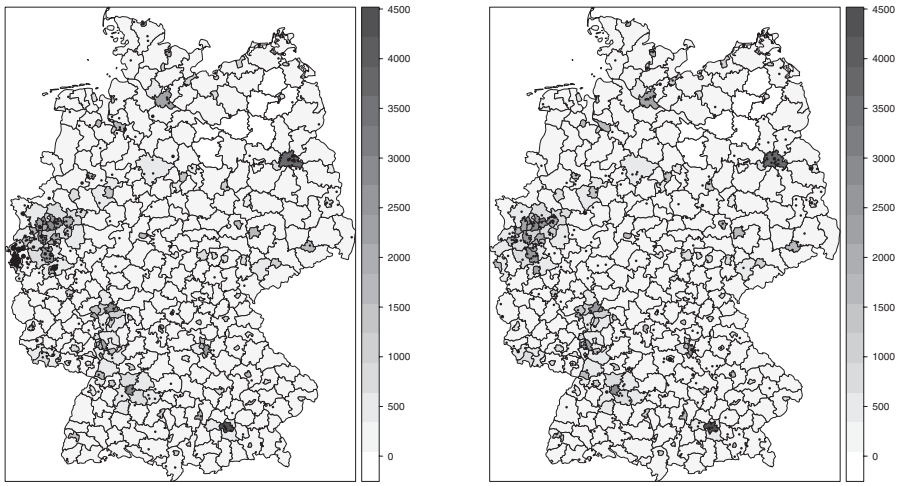
All in all, $n = 636$ infections with finetypes B (336) and C (300) have been registered. Figure 1 shows the monthly numbers of IMD cases for each finetype. Cases of IMD predominantly occur during winter and early spring, which can be seen from more or less pronounced peaks in the figure. Specifically, a connection between outbreaks of meningococci and influenza is hypothesized. For example, Jensen et al. (2004) found an association between the influenza detection rate and the number of IMD cases during the same week in temporal analysis of data from Northern Jutland County in Denmark, during 1980–1999.

Figure 2 presents the spatial distributions of the two finetypes based on the postcodes of the patients' residences. Over the 7 year period some cases shared the same postal code; therefore, the area of each point in the figure is drawn proportional to the number of cases at its location. For the serogroup B finetype in (a) the highest point multiplicity is 16, whereas for the serogroup C finetype in (b) this number is 4. In connection with the temporal occurrence of the events shown in Figure 1, the spatial distribution suggests that IMD is an endemic disease, i.e., cases can occur at any time and at any location. The maps also show the population densities of the districts, which can be assumed to be roughly proportional to the population at risk of infection. Spatial heterogeneity of the observed point patterns thus partially arises from spatial variation in the population density. Not surprisingly, the



(a) Finetype B:P1.7-2,4:F1-5. (b) Finetype C:P1.5,2:F3-3.

Figure 1. Monthly numbers of IMD cases for both finetypes separately.



(a) Finetype B:P1.7-2,4:F1-5. (b) Finetype C:P1.5,2:F3-3.

Figure 2. Spatial point patterns of the cases of meningococci by finetype during the years 2002–2008. The area of each dot is proportional to the number of cases at its location. Also shown are the population densities (inhabitants per km^2) of Germany's districts (source: Federal Statistical Office (DESTATIS) (2009)).

intensity of points in metropolitan areas like Berlin, Munich, or the Ruhr is higher. Animated graphics of the space–time locations of infections give more insight into the epidemic character of the finetypes, and can be found as Web Animation 1. Here, it appears as if finetype B exhibits a more stationary pattern than finetype C—in the sense that infections cluster more in space and time. It is supposed, yet not proven, that this phenomenon is due to differences in the mucosal immune reaction elicited; specifically, finetype B might be more successful than C in evading mucosal clearance.

Quantifying the dynamics of IMD would be an important step in the finetype characterization of IMD. We want to perform such an investigation in a spatiotemporal manner and therefore use spatiotemporal point processes as

modeling framework. Specifically, we want to establish a regression framework allowing us to quantify the transmission dynamics of IMD and its dependency on covariates. Point process modeling has in the context of epidemics been used in a discrete spatial setting in, e.g., Neal and Roberts (2004), Diggle (2006), Scheel et al. (2007), and Jewell et al. (2009). Spatiotemporal epidemic modeling in an explicit continuous spatial setting, however, is rare with Diggle, Rowlingson, and Su (2005) being one of the few examples of covariate adjusted modeling. One explanation is the balancing between optimal spatial resolution of the data and confidentiality of cases.

Recently, there have been suggestions for splitting the dynamics of infectious diseases into endemic and epidemic components; see Held, Höhle, and Hofmann (2005) for a discrete spatial–discrete time perspective and Höhle (2009) for a discrete spatial–continuous time perspective. For the continuous spatial–continuous time setting, similar modeling approaches have been seen in the analysis of earthquake data, see, e.g., Ogata (1998, 1999). Other areas of application are the modeling of forest fires (Peng, Schoenberg, and Woods, 2005), residential burglaries (Mohler et al., 2011), and the analysis of bird nesting patterns (Diggle, Kaimi, and Abellana, 2009). Altogether, our proposed modeling provides a unifying regression framework—beyond epidemics—for the modeling, inference, and simulation of spatiotemporal point processes.

This article is organized as follows: Section 2 presents the spatiotemporal two-component epidemic model based on the CIF, whereas Sections 3 and 4 discuss inference and simulation for the proposed model. Section 5 analyses the IMD data, and a discussion in Section 6 finalizes the article.

2. Spatiotemporal Two-Component CIF Model

In the following text, we propose a novel additive-multiplicative model for the CIF of an infectious disease process continuous in space–time with events occurring in a prespecified observation period $[0,\ T]$, $T > 0$, and observation region $W \subset \mathbb{R}^2$. The CIF $\lambda^*(t, \boldsymbol{s})$ represents the instantaneous rate or hazard for events at time t and location $\boldsymbol{s}$ given all the observations up to time t (the asterisk notation shall represent the conditioning on the random past history of the process).

The basic framework of the proposed model is to superimpose endemic and epidemic components to model the IMD surveillance data—an idea similar to the two-component spatial susceptible-infectious-recovered (SIR) model (Höhle, 2009):

$$\lambda^*(t, \boldsymbol{s}) = h(t, \boldsymbol{s}) + e^*(t, \boldsymbol{s}) \qquad (t > 0, \boldsymbol{s} \in W).$$

The epidemic component $e^*(t, \boldsymbol{s})$ represents the spread of the disease by person-to-person contact. The endemic component $h(t, \boldsymbol{s})$ models otherwise imported cases and is—contrary to the epidemic component—independent of the internal history of the process.

2.1 *Specification of the Endemic Component* $h(t,\ s)$

The endemic component is of the multiplicative form $h(t, \boldsymbol{s}) = \rho(t, \boldsymbol{s}) \exp(\boldsymbol{\beta}' \boldsymbol{z}(t, \boldsymbol{s}))$, where $\rho(t, \boldsymbol{s})$ is a known spatiotemporal intensity offset, e.g., the population density at time t in the district containing the location $\boldsymbol{s}$, such that the endemic rate of infection is proportional to the population density. Furthermore, $\boldsymbol{z}(t, \boldsymbol{s})$ is a linear predictor of endemic covariates, e.g., this could be a temporal trend or exogenous covariates resulting from another jointly evolving point process. For example, in the IMD application, an endemic covariate is the number of influenza cases on a week × district grid (possibly time lagged). Altogether, the endemic component is modeled as a piecewise constant function on some spatiotemporal grid resulting from a decomposition of the time period $(0,\ T]$ and the observation region W. The consecutive time intervals of this decomposition (e.g., weeks) are denoted by $C_1, \ldots,\ C_D \subset (0,\ T]$, and the spatial tiles (e.g., districts) are denoted by $A_1, \ldots,\ A_M \subset W$. Let the functions $\tau(t)$ and $\xi(\boldsymbol{s})$ return the indices of the temporal and spatial grid units containing time point t and coordinate $\boldsymbol{s}$, respectively. Then, the endemic component can be written as

$$h(t, \boldsymbol{s}) = \rho_{\tau(t),\xi(\boldsymbol{s})} \exp(\boldsymbol{\beta}' \boldsymbol{z}_{\tau(t),\xi(\boldsymbol{s})}), \tag{1}$$

where $\rho_{\tau(t),\xi(\boldsymbol{s})}$ is the known interval- and tile-specific offset and $\{\boldsymbol{z}_{\tau,\xi} : \tau \in \{1, \ldots, D\},\ \xi \in \{1, \ldots, M\}\}$ is a collection of covariates on the spatiotemporal grid $\{C_1, \ldots,\ C_D\} \times \{A_1, \ldots,\ A_M\}$.

2.2 *Specification of the Epidemic Component* $e^*(t,\ \boldsymbol{s})$

The self-exciting component of the model essentially provides a description of the infection pressure at a space–time location $(t, \boldsymbol{s})$ caused by each infectious individual. This infectivity of an infectious individual j, denoted by $e_j(t, \boldsymbol{s})$, corresponds to the inhomogeneous rate of a Poisson process, the realizations of which are the space–time locations of infected individuals. This so called triggering function is factorized into separate effects of marks, elapsed time, and relative location:

$$e_j(t, \boldsymbol{s}) = e^{\eta_j}\, g(t - t_j)\, f(\boldsymbol{s} - \boldsymbol{s}_j), \qquad (t > t_j), \tag{2}$$

where $(t_j, \boldsymbol{s}_j)$ is the infection time and location of individual j, $\eta_j = \gamma_0 + \boldsymbol{\gamma}' \boldsymbol{m}_j$ is a linear predictor based on the vector of unpredictable marks $\boldsymbol{m}_j$ attached to the infected individual, and g and f are positive temporal and spatial interaction functions, respectively. The effects $\boldsymbol{\gamma}$ of marks reflect that different individuals might cause more or less secondary cases, depending on individual characteristics.

The interaction functions describe the decay of infectivity with an increasing spatial or temporal distance from the infection source. In infectious disease applications, f is often taken to be a radially symmetric kernel corresponding to an isotropic spread of the disease, such that $f(\boldsymbol{s} - \boldsymbol{s}_j) \equiv f(||\boldsymbol{s} - \boldsymbol{s}_j||)$. A typical example is to let f be the kernel of a bivariate normal density with zero mean and diagonal covariance matrix. The temporal interaction function could be chosen as $g(t) = e^{-\alpha t}$, $t > 0$, $\alpha > 0$, representing an exponential temporal decay of infectivity (Hawkes, 1971).

The resulting epidemic component $e^*(t, \boldsymbol{s})$ is the sum of the contributions (2) of all infectious individuals at time t and location $\boldsymbol{s}$. Formally,

$$\begin{aligned} e^*(t, \boldsymbol{s}) &= \int_{(0,t)\times W \times \mathcal{M}} \mathbb{1}_{(0,\varepsilon]}(t - \tilde{t})\, \mathbb{1}_{[0,\delta]}(||\boldsymbol{s} - \tilde{\boldsymbol{s}}||)\, e^{\eta_j}\, g(t - \tilde{t}) \\ &\qquad \times f(\boldsymbol{s} - \tilde{\boldsymbol{s}})\, N(\mathrm{d}\tilde{t} \times \mathrm{d}\tilde{\boldsymbol{s}} \times \mathrm{d}\tilde{\boldsymbol{m}}), \\ &= \sum_{j \in I^*(t,\boldsymbol{s})} e^{\eta_j}\, g(t - t_j)\, f(\boldsymbol{s} - \boldsymbol{s}_j), \end{aligned} \tag{3}$$

where $\mathcal{M}$ is the mark space, N is the time–space–mark point process counting the infections and $I^*(t,\boldsymbol{s}) := \{j \in \{1,\dots,N_g(t-)\} : \mathbb{1}_{(0,\varepsilon]}(t-t_j)=1 \ \wedge\ \mathbb{1}_{[0,\delta]}(||\boldsymbol{s}-\boldsymbol{s}_j||)=1\}$ is the history-dependent set of infectives at time t and location $\boldsymbol{s}$, where $N_g(t-) = N((0,t)\times W\times\mathcal{M})$. In the above, the hyperparameters ε, $\delta > 0$ are introduced as known *maximum* temporal and spatial interaction ranges. A past event only influences the process at time t and location $\boldsymbol{s}$, if both indicator functions are true, i.e., if it occurred at most ε time units ago at a location within distance δ.

2.3 *Characteristics of the Model*

Altogether, the proposed CIF model for a self-exciting spatiotemporal point process with components (1) and (3) is

$$\lambda^*(t,\boldsymbol{s}) = \rho_{\tau(t),\xi(\boldsymbol{s})}\exp(\boldsymbol{\beta}'\boldsymbol{z}_{\tau(t),\xi(\boldsymbol{s})}) + \sum_{j\in I^*(t,\boldsymbol{s})} e^{\eta_j}\, g(t-t_j)\, f(\boldsymbol{s}-\boldsymbol{s}_j),$$

which we shall call `twinstim` to indicate a two-component spatiotemporal (conditional) intensity model. For the proposed model an interesting quantity is the individual-specific mean number μ_j of infections caused by individual j inside its spatiotemporal range of interaction:

$$\begin{aligned}\mu_j &= \int_0^\infty \int_{\mathbb{R}^2} e_j(t,\boldsymbol{s})\,\mathbb{1}_{(0,\varepsilon]}(t-t_j)\,\mathbb{1}_{[0,\delta]}(||\boldsymbol{s}-\boldsymbol{s}_j||)\,\mathrm{d}t\,\mathrm{d}\boldsymbol{s} \\ &= e^{\eta_j}\cdot\int_0^\varepsilon g(t)\,\mathrm{d}t\cdot\int_{b(\mathbf{0},\delta)} f(\boldsymbol{s})\,\mathrm{d}\boldsymbol{s}. \end{aligned} \tag{4}$$

Here, $b(\mathbf{0},\delta)$ denotes the disc centered at (0,0)' with radius δ. The integration domain $\mathbb{R}_+\times\mathbb{R}^2$ above stems from the theoretical point of view that the point process occurs in unlimited time and space. In practice this is not observable, but individuals near the border would be attributed a truncated value of μ_j if integrating over W—or, similarly, $[0,\ T]$—only. Such edge effects are overcome by (4), which also simplifies interpretation by providing a quantity similar to the basic reproduction number R_0 known from classical epidemic modeling. Specifically, the number μ_j offers an intuitive way of interpreting the parameters $\boldsymbol{\gamma}$ in the linear predictor η_j, because they can be handled as usual in Poisson regression models: a unit positive change in a specific continuous mark m_{jl} multiplies the mean number of infections by the corresponding parameter e^{γ_l}.

2.4 *Extension: Type-Specific* `twinstim`

Although the model of the previous subsection allows for a finetype-specific infectivity through the vector of unpredictable marks $\boldsymbol{m}_j$, it is not applicable for a joint modeling of both finetypes. This is because finetypes do not change during transmission. Therefore, the point process model will be extended to a marked version suitable for the specific application of IMD and point patterns with different event types in general.

Denote by $\mathcal{K} = \{1,\dots,K\} \subset \mathbb{N}$ the set of possible event types. Define an indicator matrix $\boldsymbol{Q} = (q_{k,l})_{k,l\in\mathcal{K}}$, where $q_{k,l} \in \{0;\,1\}$, which determines the possible ways of transmission. If $q_{k,l}$ equals 1, an infective type k event can cause an event of type l. For instance, the IMD data would require $\boldsymbol{Q} = \boldsymbol{I}_2$, because the transmission is finetype specific. A marked spatiotemporal point process on $(0,T]\times W\times\mathcal{K}$ is then defined by the following model for the CIF:

$$\begin{aligned}\lambda^*(t,\boldsymbol{s},\kappa) &= h(t,\boldsymbol{s},\kappa) + e^*(t,\boldsymbol{s},\kappa),\\ h(t,\boldsymbol{s},\kappa) &= \rho_{\tau(t),\xi(\boldsymbol{s})}\exp(\beta_0(\kappa)+\boldsymbol{\beta}'\boldsymbol{z}_{\tau(t),\xi(\boldsymbol{s})}),\\ e^*(t,\boldsymbol{s},\kappa) &= \sum_{j\in I^*(t,\boldsymbol{s},\kappa)} e_j(t,\boldsymbol{s}),\\ e_j(t,\boldsymbol{s}) &= \exp(\eta_j)\cdot g(t-t_j|\kappa_j)\cdot f(\boldsymbol{s}-\boldsymbol{s}_j|\kappa_j),\\ I^*(t,\boldsymbol{s},\kappa) &= \big\{j\in\{1,\dots,N_g(t-)\}:\ \mathbb{1}_{(0,\varepsilon]}(t-t_j)\\ &\qquad = 1\ \wedge\ \mathbb{1}_{[0,\delta]}(||\boldsymbol{s}-\boldsymbol{s}_j||)=1\ \wedge\ q_{\kappa_j,\kappa}=1\big\}. \end{aligned} \tag{5}$$

Here, the transmission indicators from the matrix $\boldsymbol{Q}$ have been integrated into $I^*(t,\boldsymbol{s},\kappa)$. Note that the event type κ_j is now part of the vector $\boldsymbol{m}_j$, which enables type-specific epidemic intercepts as well as type interactions with individual covariates in the linear predictor η_j. The new endemic intercept $\beta_0(\kappa)$ either represents a type-specific endemic intercept, i.e., $\beta_0(\kappa) = \sum_{k=1}^K \beta_{0,k}\,\mathbb{1}_{\{k=\kappa\}}(\kappa) = \beta_{0,\kappa}$, or contains only a single global intercept $\beta_0(\kappa) = \beta_0$, corresponding to the hypothesis $\beta_0 = \beta_{0,1} = \cdots = \beta_{0,K}$. For the remainder of the endemic predictor, the model assumes independence of κ, which means that the effect of endemic covariates is homogeneous over the event types. However, the history-dependent set $I^*(t,\boldsymbol{s},\kappa)$ of infective individuals now accounts for the transmission regime $\boldsymbol{Q}$ between the event types, and the interaction functions are allowed to depend on the type of the infective event as well.

3. Statistical Inference

This section deals with likelihood inference for the parameters of the CIF in (5) based on the observed marked spatiotemporal point pattern $\boldsymbol{x} = \{(t_i,\boldsymbol{s}_i,\boldsymbol{m}_i) : i = 1,\dots,n\}$, where the event type κ_i is part of the vector of marks $\boldsymbol{m}_i$, and n is the number of events, i.e., a realization of $N_g(T)$. The parameter vector in question is $\boldsymbol{\theta} = (\boldsymbol{\beta_0}', \boldsymbol{\beta}', \boldsymbol{\gamma}', \boldsymbol{\sigma}', \boldsymbol{\alpha}')'$, where $\boldsymbol{\sigma}$ and $\boldsymbol{\alpha}$ are the parameter vectors of the spatial and temporal interaction functions $f_{\boldsymbol{\sigma}}$ and $g_{\boldsymbol{\alpha}}$, respectively.

In our framework, no attempt is made to model unpredictable marks like gender and age but they are taken as given predictor variables in models of the CIF. In this case, the log likelihood of the underlying point process N on $[0;T]\times W\times\mathcal{M}$ may be conveniently written as (Daley and Vere-Jones, 2003)

$$\sum_{i=1}^n \log\lambda^*_{\boldsymbol{\theta}}(t_i,\boldsymbol{s}_i,\kappa_i) - \int_0^T\int_W \sum_{\kappa\in\mathcal{K}} \lambda^*_{\boldsymbol{\theta}}(t,\boldsymbol{s},\kappa)\,\mathrm{d}t\,\mathrm{d}\boldsymbol{s}.$$

The components of the above sum can be directly calculated for a specific value of the parameter vector $\boldsymbol{\theta}$ after having determined the set $I^*(t_i,\boldsymbol{s}_i,\kappa_i)$ of potential sources of infection for the ith event. Furthermore, the integrations of the endemic and epidemic components of the CIF can be performed separately due to their additive superposition. Recalling that the endemic component is a piecewise constant function on the spatiotemporal grid $\{C_1,\dots,C_D\}\times\{A_1,\dots,A_M\}$, its integral is in fact a sum over this grid of smallest observed units

in space–time:

$$\int_0^T \int_W \sum_{\kappa\in\mathcal{K}} h_{\boldsymbol{\theta}}(t,\boldsymbol{s},\kappa)\,\mathrm{d}t\,\mathrm{d}\boldsymbol{s} = \left(\sum_{\kappa\in\mathcal{K}} \exp\left(\beta_0(\kappa)\right)\right) \cdot \sum_{\tau=1}^{D}\sum_{\xi=1}^{M} |C_\tau|\,|A_\xi|\,\rho_{\tau,\xi}\,\exp\left(\boldsymbol{\beta}'\boldsymbol{z}_{\tau,\xi}\right). \tag{6}$$

The integrated epidemic component can be simplified by moving the indicators of the function $I^*(t,\boldsymbol{s},\kappa)$ back into the sum:

$$\begin{aligned}
&\int_0^T \int_W \sum_{\kappa\in\mathcal{K}} e^*_{\boldsymbol{\theta}}(t,\boldsymbol{s},\kappa)\,\mathrm{d}t\,\mathrm{d}\boldsymbol{s} \\
&= \int_0^T \int_W \sum_{\kappa\in\mathcal{K}} \sum_{j=1}^{n} \mathbb{1}_{(0,\varepsilon]}(t-t_j)\,\mathbb{1}_{[0,\delta]} \\
&\quad \times (||\boldsymbol{s}-\boldsymbol{s}_j||)\,q_{\kappa_j,\kappa}\,e^{\eta_j}\,g_{\boldsymbol{\alpha}}(t-t_j|\kappa_j) f_{\boldsymbol{\sigma}}(\boldsymbol{s}-\boldsymbol{s}_j|\kappa_j)\,\mathrm{d}t\,\mathrm{d}\boldsymbol{s} \\
&= \sum_{j=1}^{n} q_{\kappa_j,\cdot}\,e^{\eta_j} \left(\int_0^{\min\{T-t_j;\varepsilon\}} g_{\boldsymbol{\alpha}}(t|\kappa_j)\,\mathrm{d}t\right) \\
&\quad \times \left(\int_{R_j} f_{\boldsymbol{\sigma}}(\boldsymbol{s}|\kappa_j)\,\mathrm{d}\boldsymbol{s}\right).
\end{aligned} \tag{7}$$

Here, $q_{\kappa_j,\cdot} := \sum_{\kappa\in\mathcal{K}} q_{\kappa_j,\kappa}$ is the number of different event types that can be triggered by an event of type κ_j, and $R_j := \left\{W \cap b(\boldsymbol{s}_j;\delta)\right\} - \boldsymbol{s}_j$ is the spatial interaction region of the jth event centered at its location.

The evaluation of the two-dimensional integral over the domains R_j is the most sophisticated task of the log-likelihood evaluation. Meyer (2009) compared accuracy and speed of different cubature rules for performing the numerical integration. Here, the two-dimensional midpoint rule (see, e.g., Stroud, 1971) proved to be best suited for the task. In contrast, the evaluation of the definite integral over the temporal interaction function is analytically accessible for typical choices of $g_{\boldsymbol{\alpha}}$.

Altogether, an analytical maximization of the above log-likelihood is not feasible, and a numerical optimization routine such as the Broyden-Fletcher-Goldfarb-Shanno (BFGS) method (see, e.g., Nocedal and Wright, 1999, Section 8.1) is required. Here, it is advantageous to know the score function $s(\boldsymbol{\theta})$, which is derived in Web Appendix A. Uncertainty of the parameter estimates is deduced from the expected Fisher information $\mathcal{I}(\boldsymbol{\theta})$ as estimated by the "optional variation process" adapted to the marked spatiotemporal setting—see Web Appendix B for details. Significance of specific model parameters can be investigated by Wald or likelihood ratio tests and model selection is performed based on Akaike's information criterion (AIC).

4. Simulation Algorithm

In general, the usability of a model class is greatly improved by the ability to simulate from a specific model. For instance, it enables model checking and parametric bootstrap. For evolutionary point processes specified by their CIF, *Ogata's modified thinning algorithm* (Daley and Vere-Jones, 2003, Algorithm 7.5.V.) provides a convenient and exact way to simulate realizations of the process. The algorithm requires piecewise upper bounds for the intensity $\lambda^*_g(t)$ of the ground process $N_g(t) := N((0,t]\times W\times\mathcal{K})$. This intensity is determined as

$$\begin{aligned}
\lambda^*_g(t) = \int_W \sum_{\kappa\in\mathcal{K}} \lambda^*(t,\boldsymbol{s},\kappa)\,\mathrm{d}\boldsymbol{s} &= \left(\sum_{\kappa\in\mathcal{K}} e^{\beta_0(\kappa)}\right) \\
&\times \left(\sum_{\xi=1}^{M} |A_\xi|\,\rho_{\tau(t),\xi}\,e^{\boldsymbol{\beta}'\boldsymbol{z}_{\tau(t),\xi}}\right) \\
&+ \sum_{j=1}^{N_g(t-)} \left(\sum_{\kappa\in\mathcal{K}} q_{\kappa_j,\kappa}\right) e^{\eta_j}\,\mathbb{1}_{(0,\varepsilon]}(t-t_j) \\
&\times g(t-t_j|\kappa_j) \int_{R_j} f(\boldsymbol{s}|\kappa_j)\,\mathrm{d}\boldsymbol{s}.
\end{aligned}$$

This function is bounded above by the CIF $\overline{\lambda^*_g}(t)$, which is defined by replacing $g(t|\kappa)$ by the *constant* temporal interaction function $\overline{g}(t|\kappa) = \max_{u>0} g(u|\kappa)$. This CIF is piecewise constant in time as it only jumps at time points where any of the endemic covariates in $\boldsymbol{z}_{\tau(t),\xi}$ in any tile ξ changes its value, or when the set of currently infectious individuals changes, i.e., whenever a new event occurs or a previous event stops triggering.

Given a parameter vector $\boldsymbol{\theta}$, the ranges of interaction ε and δ, as well as a sampling scheme for the marks $\boldsymbol{m}_j$, the time point of the next infection starting from the current time $t = t_0$ can be generated as follows: Draw an exponentially distributed random variate Δ with rate $\overline{\lambda^*_g}(t_0)$. The simulated value of Δ is a proposal for the waiting time to the next event, i.e., the next time point of infection might be $\tilde{t} = t_0 + \Delta$. However, this proposal is not valid if the rate $\overline{\lambda^*_g}(t)$ had changed between t_0 and $\tilde{t}$. In this case, time is set to the first changepoint after t_0 and a new Δ is simulated. Eventually, a proposed time point $\tilde{t}$ is valid. It is then accepted with probability $\lambda^*_g(\tilde{t})/\overline{\lambda^*_g}(\tilde{t})$. If it is rejected, time is set to $t = \tilde{t}$ and a new waiting time Δ is simulated as above. If it is accepted, location $\tilde{\boldsymbol{s}}$ and type $\tilde{\kappa}$ of the event have to be simulated. At first, the source of infection is sampled with probabilities proportional to the respective components of $\lambda^*_g(\tilde{t})$:

$$\begin{aligned}
&\mathbb{P}(\text{endemic source})\cdot\lambda^*_g(\tilde{t}) \\
&\quad = \left(\sum_{\kappa\in\mathcal{K}} e^{\beta_0(\kappa)}\right)\left(\sum_{\xi=1}^{M} |A_\xi|\,\rho_{\tau(\tilde{t}),\xi}\,e^{\boldsymbol{\beta}'\boldsymbol{z}_{\tau(\tilde{t}),\xi}}\right), \\
&\mathbb{P}(\text{source} = \text{event } j)\cdot\lambda^*_g(\tilde{t}) \\
&\quad = \left(\sum_{\kappa\in\mathcal{K}} q_{\kappa_j,\kappa}\right) e^{\eta_j}\,\mathbb{1}_{(0,\varepsilon]}(\tilde{t}-t_j)\,g(\tilde{t}-t_j|\kappa_j) \int_{R_j} f(\boldsymbol{s}|\kappa_j)\,\mathrm{d}\boldsymbol{s},
\end{aligned} \tag{8}$$

for $j\in\{1,\ldots,N_g(\tilde{t}-)\}$. On the one hand, if the new event has an endemic source, then $\mathbb{P}(\tilde{\kappa}=k)\propto\exp(\beta_0(k))$, $k\in\mathcal{K}$, and $\mathbb{P}(\tilde{\boldsymbol{s}}\in A_\xi)\propto|A_\xi|\,\rho_{\tau(\tilde{t}),\xi}\,e^{\boldsymbol{\beta}'\boldsymbol{z}_{\tau(\tilde{t}),\xi}}$, $\xi=1,\ldots,M$. In the sampled tile $A_{\tilde{\xi}}$, the location $\tilde{\boldsymbol{s}}$ is uniformly distributed. On the other hand, if the new event was triggered by the previous event j, then $\tilde{\kappa}\sim U(\{k: q_{\kappa_j,k}=1\})$, and $\tilde{\boldsymbol{s}} = \boldsymbol{s}_j + \boldsymbol{v}$, where $\boldsymbol{v}$

is drawn from the density $f(\boldsymbol{s}|\kappa_j)/\int_{R_j} f(\boldsymbol{s}|\kappa_j)\,\mathrm{d}s$ on R_j, e.g., using rejection sampling.

A scheme of the described algorithm can be found as Web Appendix C.

5. Application to the IMD Data

Although visual comparisons between the finetypes and heuristic comparisons of the estimates of separate finetype-specific models are possible, this does not allow to assess potential differences statistically. We thus conduct a joint analysis of the two finetypes by the marked `twinstim` of Section 2.4. We perform model selection for the joint point pattern of 630 cases of IMD with complete age and gender information by using AIC to compare all models with the CIF composed by subsets of the following terms:

- Endemic component: common or finetype-specific intercept, linear time trend, time-of-year effects (one or two harmonics), and linear effect of weekly number of influenza cases registered in the district of a point (no time lag, lags 0 and 1, lags 0–2, or lags 0–3) taken from the SurvStat database (Robert Koch-Institut, 2009).
- Epidemic component: gender, age (categorized as 0–2, 3–18, and ≥19 years), finetype, and age–finetype interaction.

As an offset in the endemic component, we use the district-specific population density $\rho_{\xi(\boldsymbol{s})}$ (inhabitants per km^2). A fixed hyperparameter of $\varepsilon = 30$ days is assumed—this maximal temporal interaction range is consistent with the range used in, e.g., Zangwill et al. (1997). Because the number of supposedly direct transmissions in the IMD dataset is humble, we will furthermore assume a constant temporal interaction function g (i.e., constant spread within the ε days) to not overparametrize the epidemic component. The spatial hyperparameter is fixed at $\delta = 200$ km—this parameter needs only to be large enough not to influence the estimation of the actual spatial interaction function f.

To restrict the model search, and hence computing time, we first performed the search for all 600 models ($2\cdot2\cdot3\cdot5$ configurations of the endemic component and $2\cdot5$ configurations of the epidemic component) with constant spatial interaction function f. Hereafter, the top 10 models of this search were investigated further with two Gaussian spatial interaction functions: one with joint variance parameter and one with finetype-specific variance parameter.

The CIF of the resulting AIC-best model obtained by this search was $\lambda^*_{\boldsymbol{\theta}}(t,\boldsymbol{s},\kappa) =$

$$\begin{aligned}
&\rho_{\xi(\boldsymbol{s})} \cdot \exp\left(\beta_0 + \beta_{\text{trend}}\frac{\lfloor t\rfloor}{365} + \beta_{\sin}\sin\left(\lfloor t\rfloor\,\frac{2\pi}{365}\right)\right.\\
&\qquad\left. + \beta_{\cos}\cos\left(\lfloor t\rfloor\,\frac{2\pi}{365}\right)\right)\\
&+ \sum_{j\in I^*(t,\boldsymbol{s},\kappa)} q_{\kappa_j,\kappa}\ \exp(\gamma_0 + \gamma_{3\text{-}18}\ \mathbb{1}_{[3,18]}(\text{age}_j)\\
&\qquad + \gamma_{\geq 19}\ \mathbb{1}_{[19,\infty)}(\text{age}_j) + \gamma_{\text{C}}\ \mathbb{1}_{\{C\}}(\kappa_j))\ f_\sigma(\boldsymbol{s}-\boldsymbol{s}_j).
\end{aligned}$$

Here, $(t,\boldsymbol{s},\kappa)$ denotes days since December 31, 2001, coordinate in ETRS89 (kilometer scale), and finetype. With $\lfloor t\rfloor$ we denote Monday of week $\tau(t)$, i.e., the lower bound of time intervals $C_1,\ldots,C_D$. In the linear predictor of the epidemic component, age group 0-2 and type B serve as reference categories. The corresponding parameter estimates of the best model, now fitted to the 635 cases with available age, are found in Table 1.

Table 1
Parameter estimates of the model with the lowest AIC (AIC=18968). The p-values correspond to Wald tests.

	Estimate	Std. error	z-value	$\mathbb{P}(\lvert Z\rvert > \lvert z\rvert)$
β_0	−20.3652	0.0872	−233.53	$<2\cdot10^{-16}$
β_{trend}	−0.0493	0.0223	−2.21	0.027
$\beta_{\sin}$	0.2618	0.0649	4.03	$5.5\cdot10^{-05}$
$\beta_{\cos}$	0.2668	0.0644	4.14	$3.4\cdot10^{-05}$
γ_0	−12.5746	0.3128	−40.21	$<2\cdot10^{-16}$
$\gamma_{3\text{-}18}$	0.6463	0.3195	2.02	0.04310
$\gamma_{\geq 19}$	−0.1868	0.4321	−0.43	0.66558
γ_{C}	−0.8496	0.2574	−3.30	0.00097
$\log\sigma$	2.8287	0.0819		

Thus, there appears to be no noteworthy difference in the endemic behavior of the two types: a linear downward time trend superimposed with one harmonic best describes the endemic behavior of the point pattern (see Figure 3a). An additional effect of past numbers of influenza cases does not improve the model. In contrast, there is an effect of past IMD cases, i.e., the process is indeed self-exciting. Comparing the endemic-only model with the model enriched by an epidemic intercept only, greatly improves the fit (ΔAIC=202.84). In the epidemic component, there is a detectable dependence on marks with type C being less aggressive than type B. Figure 3b shows the resulting finetype-specific spatial interaction functions, which for type C is $e^{\hat\gamma_C}\cdot 100\% = 43\%$ of type B. Finally, there is a significant age difference in the infectivity of cases: the highest potential is found in the 3–18 year old, which could be interpreted as the kindergarten and school-aged children having a higher contact behavior than, e.g., adults.

Based on the selected model, basic reproduction numbers of $\hat\mu_{\text{B}} = 0.25$ (95% CI 0.19–0.34) versus $\hat\mu_{\text{C}} = 0.11$ (95% CI 0.07–0.17) are obtained by calculating the type-specific expectation of (4) over the empirical distribution function of the additional covariates in the epidemic predictor (here: age group). The confidence intervals are given as the 0.025 and 0.975 quantiles of samples obtained by recomputing $\hat\mu_{\text{B}}$ and $\hat\mu_{\text{C}}$ for 999 additional coefficient vectors drawn from the asymptotic multivariate normal distribution of the parameter estimates in Table 1. The confidence intervals thus indicate a higher epidemic potential of the serogroup B finetype. Note that these numbers are lower than what one would expect from the literature, e.g., Trotter, Gay, and Edmunds (2005) report an R_0 estimate of 1.36 for *serogroup* C. Two explanations account for this discrepancy: firstly, our estimation is based on transmission between cases with invasive disease and not between asymptomatic carriers, who are not represented in disease surveillance data. Secondly, use of an endemic component means that our R_0 estimates are destined to be lower, because sporadic cases do not contribute to the number of

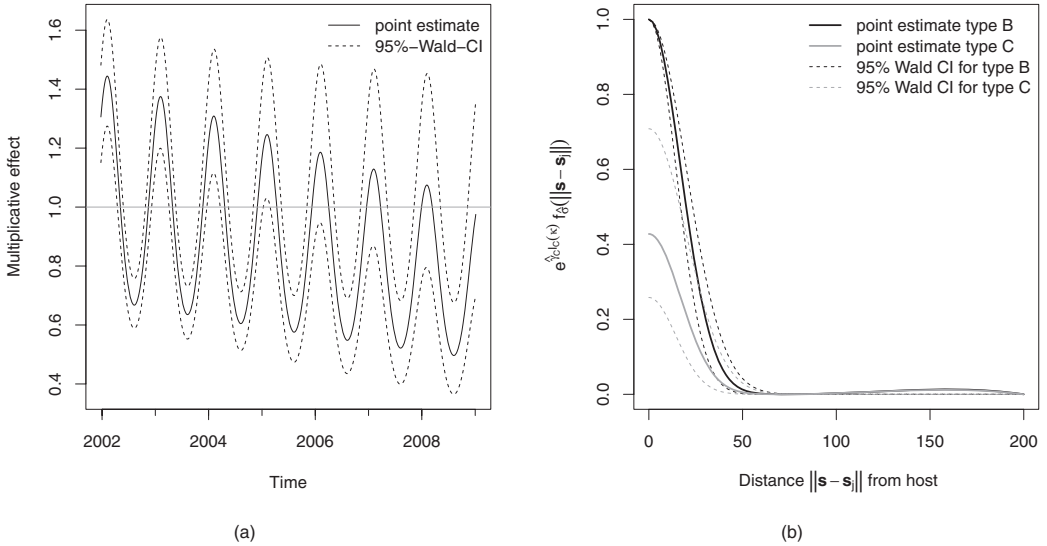


Figure 3. (a) Trend and seasonal component of the fitted model; one observes the typical IMD peak in late February and minimum in August. Furthermore, (b) shows the spatial interaction function multiplied by the type modifier illustrating the higher epidemic potential of type B.

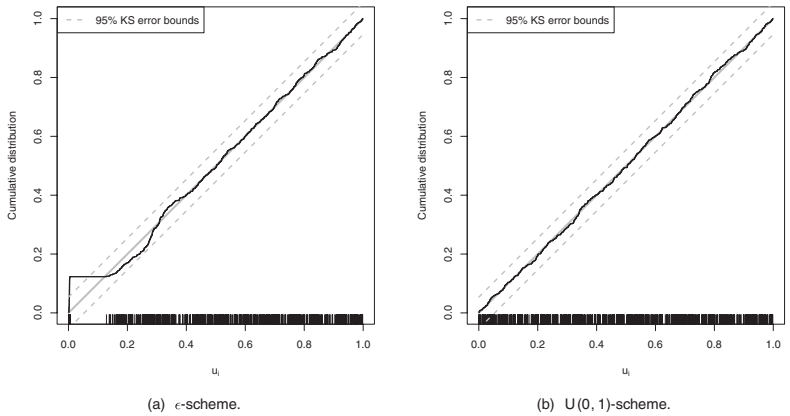


Figure 4. CDF of the observed U_i together with 95% Kolmogorov–Smirnov error bounds for data with tie breaking according to the (a) ϵ scheme and (b) $U(0,\,1)$ scheme.

secondary cases. Still, our estimates provide realistic lower bounds for carriage reproduction numbers.

To inspect the goodness of fit of the selected spatiotemporal point process model, we follow the suggestion by Ogata (1988) (see also Rathbun, 1996) by computing $Y_i = \hat{\Lambda}_g^*(t_i) - \hat{\Lambda}_g^*(t_{i-1})$, $i = 2, \ldots, n$, where $\hat{\Lambda}_g^*(t)$ is the fitted cumulative intensity function of the ground process. If the estimated CIF describes the true CIF well, then $U_i = 1 - \exp(-Y_i) \overset{\text{iid}}{\sim} U(0, 1)$. Figure 4a contains a plot of the cumulative density function (CDF) of the observed U_i and for comparison the CDF of the $U(0,\,1)$-distribution together with error bounds computed by inverting the one sample Kolmogorov–Smirnov test. The fit appears good, but noticeable deviations for $u_i < 0.15$ can be observed, which we suspect to occur due to the

tie-breaking strategy of subtracting $\epsilon = 0.01$ days from ties. As observations are on a per-day basis and thus are interval censored we reestimated the model for a dataset where ties were broken by subtracting a $U(0, 1)$-distributed random number from each observation time. Figure 4b shows the improved fit of this analysis—the relative changes in the parameter estimates are minor.

Another way of assessing the goodness of fit is by simulation from the fitted CIF. Figure 5 shows the observed 7 year incidences (per 100,000 inhabitants) of the 413 districts for both finetypes together. To identify extreme observations that are not explained by the selected model, we simulated 100 realizations of the process and determined the 2.5% and 97.5% quantiles of the district-specific 7 year incidences. In the figure, districts with observed incidences outside the simulated 95%-range are marked by triangles. Many of the 17 districts with an excess are found around the city Aachen at the border to the Netherlands. The deviation from the model could thus be explained by edge effects hiding potential transmissions across the border.

Altogether, we are led to the conclusion that the proposed model provides a useful description of the spread of IMD. It allows a quantification that the serogroup B finetype has a higher epidemic potential than the serogroup C finetype and shows age difference in spread behavior. A sensitivity analysis confirmed robustness of these results for increasing values of δ. Order and significance of the finetype difference in the epidemic component remained stable for ε in the range of 1–5 weeks to 1–4 months. Age group results were slightly more varying: the 3–18 year olds remain having the highest epidemic potential, but from $\varepsilon > 35$ days on, the oldest age group comes in second. The sensitivity analysis also showed, that there is too little information to estimate ε from the IMD data—we are thus forced to fix the hyperparameter at a biological plausible value.

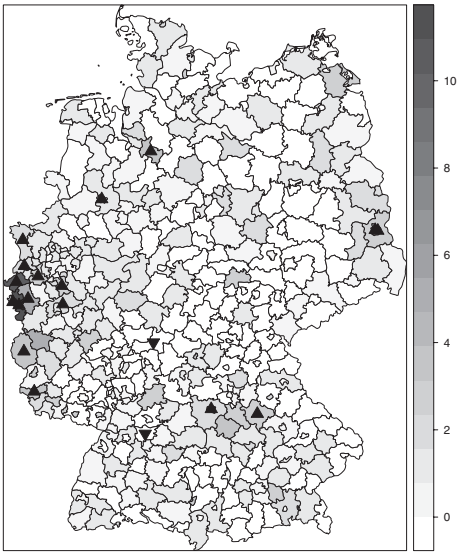


Figure 5. Observed incidence (per 100,000 inhabitants) during 2002–2008 for both finetypes together. Triangles pointing up (down) indicate districts with a higher (lower) incidence than explained by 100 simulations from the model.

6. Discussion

We presented a comprehensive framework for modeling, inference, and simulation for infectious disease occurrence data. In the case of IMD, the infected individual is effectively removed from the transmission network once the disease becomes manifest. Secondary cases are thought to acquire the infective strain either from the case during incubation or from asymptomatic carriers close to the case. Although marks attached to the case can naturally not account for the latter mode of transmission, they represent a valid proxy for the transmission network of the case when analyzing surveillance data, which typically lack information regarding carriage.

Despite use of disease surveillance data, we were able to quantify differences in IMD transmission dynamics based on age and finetype. That the modeling requires an epidemic component is of epidemiological interest in its own, as this shows that IMD incidence goes beyond sporadic occurrences. To our knowledge, our analysis is the first report of finetype-specific differences in spread tendencies. Contrary to previous analyses we were not able to find a significant connection between IMD and concurrent number of influenza cases. The spatial spread appeared to happen at a rather small scale—a scale which the usual district resolution data collected as part of the German Infection Protection Act does not allow to analyze. Thus our work is also a contribution to the controversy between patient privacy and the need for high-resolution data to gain new epidemiological insights. One important question in this debate is how good a proxy the patient's residence is for his general whereabouts.

Even though our CIF modeling is similar in form to the proposal in Höhle (2009), the *continuous space* of the IMD application makes epidemic modeling conceptually different. The classical SIR model framework does not apply in this situation, because events do not originate from a predefined population and individuals cannot be partitioned into model compartments anymore. Thus, including population density becomes important and one needs to distinguish between covariate information of events and covariates of the surrounding environment within which the process occurs. Furthermore, likelihood inference is complicated by requiring an additional integration over space for complex polygons. Finally, the now proposed space–time interaction functions are completely general in form and thus provide an advantage over the previous linear basis decomposition and resulting parameter constraints.

An issue currently not dealt with in our estimation are edge effects, i.e., data are only available for Germany, but infections occur outside the observation window. For example, Elias et al. (2010) investigate the contribution of cross-border spread to increased incidence of IMD in the

German region of Aachen neighboring the Netherlands. A cross-border effect is indeed detected by our simulation in Figure 5, where the Aachen region has higher observed incidences than can be explained by our model. Hence, the actual disease clusters are wider than observed in Germany, which potentially causes underestimation of the epidemic weight. Edge correction for inference in spatiotemporal point processes is, however, still an open methodological issue.

An additional strength of the proposed modeling is that it offers a parametric framework for conducting prospective change-point analysis in spatiotemporal point processes typical in disease surveillance: Within the framework of stochastic process control one could, e.g., use likelihood ratio detectors to monitor the time point where inclusion of an epidemic component is necessary to describe the observed data. This would correspond in idea to the time series setting investigated in Höhle and Paul (2008) or the homogeneous spatiotemporal Poisson process setting of Assunçáo and Correa (2009).

The presented methods for inference and simulation of `twinstim` models are available as part of the `R` package `surveillance` (Höhle, 2007; Höhle, Meyer, and Paul, 2011) available from the Comprehensive `R` Archive Network.

7. Supplementary Materials

The Web Animation referenced in Section 1 and the Web Appendices referenced in Sections 3 and 4 are available under the Paper Information link at the *Biometrics* website `http://www.biometrics.tibs.org`.

Acknowledgements

We thank Ludwig Fahrmeir for providing helpful suggestions and comments. Financial support was provided by the Munich Center of Health Sciences. Ulrich Vogel is thanked for his efforts in ensuring the generation of high-quality IMD surveillance data and helpful discussions. Matthias Frosch is acknowledged for continuous support. We thank the coeditor Thomas Louis, an anonymous associate editor, and two anonymous referees for their useful comments that improved the presentation of the article.

References

Assunçáo, R. and Correa, T. (2009). Surveillance to detect emerging space-time clusters. *Computational Statistics & Data Analysis* **53,** 2817–2830.

Daley, D. J. and Vere-Jones, D. (2003). *An Introduction to the Theory of Point Processes*, volume I: *Elementary Theory and Methods of Probability and its Applications*, 2nd edition. New York: Springer-Verlag.

Diggle, P. J. (2006). Spatio-temporal point processes, partial likelihood, foot and mouth disease. *Statistical Methods in Medical Research* **15,** 325–336.

Diggle, P. J., Rowlingson, B., and Su, T.-l. (2005). Point process methodology for on-line spatio-temporal disease surveillance. *Environmetrics* **16,** 423–434.

Diggle, P. J., Kaimi, I., and Abellana, R. (2009). Partial-likelihood analysis of spatio-temporal point-process data. *Biometrics* **66,** 347–354.

Elias, J., Schouls, L. M., van de Pol, I., Keijzers, W. C., Martin, D. R., Glennie, A., Oster, P., Frosch, M., Vogel, U., and van der Ende, A. (2010). Vaccine preventability of meningococcal clone, Greater Aachen Region, Germany. *Emerging Infectious Diseases* **16,** 465–472.

Federal Statistical Office (DESTATIS). (2009). Gemeindeverzeichnis GV 2000, Wiesbaden, Germany. Districts as of 31/12/2008. Data as of 31/12/2007.

Hawkes, A. G. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika* **58,** 83–90.

Held, L., Höhle, M., and Hofmann, M. (2005). A statistical framework for the analysis of multivariate infectious disease surveillance data. *Statistical Modelling* **5,** 187–199.

Höhle, M. (2007). Surveillance: an R package for the monitoring of infectious diseases. *Computational Statistics* **22,** 571–582.

Höhle, M. (2009). Additive-multiplicative regression models for spatio-temporal epidemics. *Biometrical Journal* **51,** 961–978.

Höhle, M. and Paul, M. (2008). Count data regression charts for the monitoring of surveillance time series. *Computational Statistics & Data Analysis* **52,** 4357–4368.

Höhle, M., Meyer, S., and Paul, M. (2011). *`Surveillance`: Temporal and spatio-temporal modeling and monitoring of epidemic phenomena*. R package version 1.3-1. Available at: http://cran.r-project.org/. Accessed September 30, 2011.

Jensen, E. S., Lundbye-Christensen, S., Samuelsson, S., Sørensen, H. T., and Schønheyder, H. C. (2004). A 20-year ecological study of the temporal association between influenza and meningococcal. *European Journal of Epidemiology* **19,** 181–187.

Jewell, C. P., Kypraios, T., Neal, P., and Roberts, G. O. (2009). Bayesian analysis for emerging infectious diseases. *Bayesian Analysis* **4,** 465–496.

Meyer, S. (2009). Spatio-temporal infectious disease epidemiology based on point processes. Master's Thesis, Department of Statistics, Ludwig-Maximilians-Universität, München. Available at: http://epub.ub.uni-muenchen.de/11703/. Accessed September 30, 2011.

Mohler, G. O., Short, M. B., Brantingham, P. J., Schoenberg, F. P., and Tita, G. E. (2011). Self-exciting point process modeling of crime. *Journal of the American Statistical Association* **106,** 100–108.

Neal, P. and Roberts, G. O. (2004). Statistical inference and model selection for the 1861 Hagelloch measles epidemic. *Biostatistics* **5,** 249–261.

Nocedal, J. and Wright, S. J. (1999). *Numerical Optimization*. New York: Springer.

Ogata, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical Association* **83,** 9–27.

Ogata, Y. (1998). Space-time point-process models for earthquake occurrences. *Annals of the Institute of Statistical Mathematics* **50,** 379–402.

Ogata, Y. (1999). Seismicity analysis through point-process modeling: a review. *Pure and Applied Geophysics* **155,** 471–507.

Peng, R. D., Schoenberg, F. P., and Woods, J. A. (2005). A space-time conditional intensity model for evaluating a wildfire hazard index. *Journal of the American Statistical Association* **100,** 26–35.

Rathbun, S. L. (1996). Asymptotic properties of the maximum likelihood estimator for spatio-temporal point processes. *Journal of Statistical Planning and Inference* **51,** 55–74.

Robert Koch Institut. (2009). SurvStat@RKI, Berlin. Available at: http://www3.rki.de/SurvStat. Accessed August 11, 2010.

Rosenstein, N. E., Perkins, B. A., Stephens, D. S., Popovic, T., and Hughes, J. M. (2001). Meningococcal disease. *The New England Journal of Medicine* **344,** 1378–1388.

Scheel, I., Aldrin, M., Frigessi, A., and Jansen, P. A. (2007). A stochastic model for infectious salmon anemia (ISA) in Atlantic salmon farming. *Journal of the Royal Society, Interface* **4,** 699–706.

Stroud, A. H. (1971). *Approximate Calculation of Multiple Integrals*. Englewood Cliffs, New Jersey: Prentice Hall.

Trotter, C., Gay, G. J., and Edmunds, W. J. (2005). Dynamic models of meningococcal carriage, disease, and the impact of serogroup c conjugate vaccination. *American Journal of Epidemiology* **162,** 89–100.

Zangwill, K. M., Schuchat, A., Riedo, F. X., Pinner, R. W., Koo, D. T., Reeves, M. W., and Wenger, J. D. (1997). School-based clusters of meningococcal disease in the United States. *JAMA* **277,** 389–395.

Received November 2010. *Revised August* 2011. *Accepted August* 2011.

Web-based Supplementary Materials for "A space-time conditional intensity model for invasive meningococcal disease occurence"

by

Sebastian Meyer[1,2], Johannes Elias[3], and Michael Höhle[4,2]

[1]Department of Psychiatry and Psychotherapy, Ludwig-Maximilians-Universität, München, Germany

[2]Department of Statistics, Ludwig-Maximilians-Universität, München, Germany

[3]German Reference Centre for Meningococci, University of Würzburg, Würzburg, Germany

[4]Department for Infectious Disease Epidemiology, Robert Koch Institute, Berlin, Germany

August 3, 2011

Web Appendix A: Calculus of the Score Function

Let $\boldsymbol{\vartheta}$ denote any subvector of $\boldsymbol{\theta}$. Then, the partial derivative of the log-likelihood with respect to $\boldsymbol{\vartheta}$ is

$$\boldsymbol{s}_{\boldsymbol{\vartheta}}(\boldsymbol{\theta}) := \frac{\partial}{\partial \boldsymbol{\vartheta}} l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{\frac{\partial}{\partial \boldsymbol{\vartheta}} \lambda^*_{\boldsymbol{\theta}}(t_i, \boldsymbol{s}_i, \kappa_i)}{\lambda^*_{\boldsymbol{\theta}}(t_i, \boldsymbol{s}_i, \kappa_i)} - \int_0^T \int_W \sum_{\kappa \in \mathcal{K}} \frac{\partial}{\partial \boldsymbol{\vartheta}} \lambda^*_{\boldsymbol{\theta}}(t, \boldsymbol{s}, \kappa) \, \mathrm{d}t \, \mathrm{d}\boldsymbol{s} \,, \tag{1}$$

and the score function is $\boldsymbol{s}(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} l(\boldsymbol{\theta}) = \left(\boldsymbol{s}'_{\beta_0}, \boldsymbol{s}'_{\beta}, \boldsymbol{s}'_{\gamma}, \boldsymbol{s}'_{\boldsymbol{\sigma}}, \boldsymbol{s}'_{\boldsymbol{\alpha}}\right)' (\boldsymbol{\theta})$. The necessary partial derivatives of the CIF with their respective time-space-mark integrals are given in the following subsections and can be plugged into the equation (1). The analytic derivatives of the interaction function f and g with respect to $\boldsymbol{\sigma}$ and $\boldsymbol{\alpha}$, respectively, have to be determined for the specific model at hand. For instance, a type-specific spatial Gaussian kernel

$$f_{\boldsymbol{\sigma}}(\boldsymbol{s}|\kappa) = \exp\left(-\frac{\|\boldsymbol{s}\|^2}{2\sigma_\kappa^2}\right)$$

with $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_K)'$ has partial derivatives

$$\frac{\partial}{\partial \sigma_k} f_{\boldsymbol{\sigma}}(\boldsymbol{s}|\kappa) = \mathbb{1}_{k=\kappa}(\kappa) \cdot \exp\left(-\frac{\|\boldsymbol{s}\|^2}{2\sigma_k^2}\right) \frac{\|\boldsymbol{s}\|^2}{\sigma_k^3}, \qquad \text{for any } k \in \mathcal{K} \,.$$

The type-specific temporal interaction function $g_{\boldsymbol{\alpha}}(t|\kappa) = e^{-\alpha_\kappa t}$ with $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ has partial derivatives $\frac{\partial}{\partial \alpha_k} g_{\boldsymbol{\alpha}}(t|\kappa) = \mathbb{1}_{k=\kappa}(\kappa) \cdot (-t\, e^{-\alpha_k t})$, for any $k \in \mathcal{K}$. While the integral of $\frac{\partial}{\partial \sigma_\kappa} f_{\boldsymbol{\sigma}}(\boldsymbol{s}|\kappa)$ over the region R_j will be approximated by numerical integration, the temporal function $\frac{\partial}{\partial \alpha_\kappa} g_{\boldsymbol{\alpha}}(t|\kappa)$ is assumed to permit analytical integration.

Endemic intercept(s) $\boldsymbol{\beta_0}$:

Let $\beta_{0,k}$, $k \in \{1, \ldots, K\}$ be one of the type-specific intercepts in $\boldsymbol{\beta_0}$. Then,

$$\frac{\partial}{\partial \beta_{0,k}} \lambda^*_{\boldsymbol{\theta}}(t, \boldsymbol{s}, \kappa) = \mathbb{1}_{k=\kappa}(\kappa) \cdot \exp\left(\beta_{0,k} + o_{\xi(\boldsymbol{s})} + \boldsymbol{\beta}' \boldsymbol{z}_{\tau(t),\xi(\boldsymbol{s})}\right)$$

since the parameter $\beta_{0,k}$ appears in the endemic component $h_{\boldsymbol{\theta}}(t, \boldsymbol{s}, \kappa)$ if and only if $\kappa = k$. The corresponding integrated value is

$$\int_0^T \int_W \sum_{\kappa \in \mathcal{K}} \frac{\partial}{\partial \beta_{0,k}} \lambda^*_{\boldsymbol{\theta}}(t, \boldsymbol{s}, \kappa) \, \mathrm{d}t \, \mathrm{d}\boldsymbol{s} = e^{\beta_{0,k}} \cdot \sum_{\tau=1}^{D} \sum_{\xi=1}^{M} |C_\tau||A_\xi| \exp\left(o_\xi + \boldsymbol{\beta}' \boldsymbol{z}_{\tau,\xi}\right) \ ,$$

cf. the integral of the endemic component in equation (6) of the paper. If the model assumes a type-invariant endemic intercept $\boldsymbol{\beta_0} = \beta_0$, then

$$\frac{\partial}{\partial \beta_0} \lambda^*_{\boldsymbol{\theta}}(t, \boldsymbol{s}, \kappa) = \exp\left(\beta_0 + o_{\xi(\boldsymbol{s})} + \boldsymbol{\beta}' \boldsymbol{z}_{\tau(t),\xi(\boldsymbol{s})}\right)$$

with integrated value

$$\int_0^T \int_W \sum_{\kappa \in \mathcal{K}} \frac{\partial}{\partial \beta_0} \lambda^*_{\boldsymbol{\theta}}(t, \boldsymbol{s}, \kappa) \, \mathrm{d}t \, \mathrm{d}\boldsymbol{s} = K \, e^{\beta_0} \cdot \sum_{\tau=1}^{D} \sum_{\xi=1}^{M} |C_\tau||A_\xi| \exp\left(o_\xi + \boldsymbol{\beta}' \boldsymbol{z}_{\tau,\xi}\right) \ .$$

Endemic covariate effects $\boldsymbol{\beta}$:

$$\frac{\partial}{\partial \boldsymbol{\beta}} \lambda^*_{\boldsymbol{\theta}}(t, \boldsymbol{s}, \kappa) = \exp\left(\beta_0(\kappa) + o_{\xi(\boldsymbol{s})} + \boldsymbol{\beta}' \boldsymbol{z}_{\tau(t),\xi(\boldsymbol{s})}\right) \cdot \boldsymbol{z}_{\tau(t),\xi(\boldsymbol{s})}$$

with corresponding integral vector (element-wise integral values)

$$\left(\sum_{\kappa \in \mathcal{K}} \exp\left(\beta_0(\kappa)\right)\right) \cdot \sum_{\tau=1}^{D} \sum_{\xi=1}^{M} |C_\tau||A_\xi| \exp(o_\xi + \boldsymbol{\beta}' \boldsymbol{z}_{\tau,\xi}) \, \boldsymbol{z}_{\tau,\xi} \ .$$

Epidemic effects γ:

$$\frac{\partial}{\partial \boldsymbol{\gamma}} \lambda^*_{\boldsymbol{\theta}}(t, \boldsymbol{s}, \kappa) = \sum_{j \in I^*(t,\boldsymbol{s},\kappa)} e^{\boldsymbol{\gamma}' \boldsymbol{m}_j} \, g_{\boldsymbol{\alpha}}(t - t_j|\kappa_j) \, f_{\boldsymbol{\sigma}}(\boldsymbol{s} - \boldsymbol{s}_j|\kappa_j) \, \boldsymbol{m}_j \, ,$$

and the corresponding integral can be deduced similar to equation (7) of the paper as

$$\sum_{j=1}^{n} q_{\kappa_j,\bullet} \, e^{\boldsymbol{\gamma}' \boldsymbol{m}_j} \left[\int_0^{\min\{T-t_j;\varepsilon\}} g_{\boldsymbol{\alpha}}(t|\kappa_j) \, \mathrm{d}t \right] \left[\int_{R_j} f_{\boldsymbol{\sigma}}(\boldsymbol{s}|\kappa_j) \, \mathrm{d}\boldsymbol{s} \right] \boldsymbol{m}_j \, .$$

Parameters σ and α of the interaction functions:

For a general spatial kernel $f_{\boldsymbol{\sigma}}(\boldsymbol{s}|\kappa)$,

$$\frac{\partial}{\partial \boldsymbol{\sigma}} \lambda^*_{\boldsymbol{\theta}}(t, \boldsymbol{s}, \kappa) = \sum_{j \in I^*(t,\boldsymbol{s},\kappa)} e^{\eta_j} \, g_{\boldsymbol{\alpha}}(t - t_j|\kappa_j) \, \left[\frac{\partial}{\partial \boldsymbol{\sigma}} f_{\boldsymbol{\sigma}}(\boldsymbol{s} - \boldsymbol{s}_j|\kappa_j) \right]$$

with corresponding integral

$$\sum_{j=1}^{n} q_{\kappa_j,\bullet} \, e^{\eta_j} \left[\int_0^{\min\{T-t_j;\varepsilon\}} g_{\boldsymbol{\alpha}}(t|\kappa_j) \, \mathrm{d}t \right] \left[\int_{R_j} \frac{\partial}{\partial \boldsymbol{\sigma}} f_{\boldsymbol{\sigma}}(\boldsymbol{s}|\kappa_j) \, \mathrm{d}\boldsymbol{s} \right] .$$

Similarly, for a general temporal kernel $g_{\boldsymbol{\alpha}}(t|\kappa)$,

$$\frac{\partial}{\partial \boldsymbol{\alpha}} \lambda^*_{\boldsymbol{\theta}}(t, \boldsymbol{s}, \kappa) = \sum_{j \in I^*(t,\boldsymbol{s},\kappa)} e^{\eta_j} \, \left[\frac{\partial}{\partial \boldsymbol{\alpha}} g_{\boldsymbol{\alpha}}(t - t_j|\kappa_j) \right] \, f_{\boldsymbol{\sigma}}(\boldsymbol{s} - \boldsymbol{s}_j|\kappa_j)$$

with corresponding integral

$$\sum_{j=1}^{n} q_{\kappa_j,\bullet} \, e^{\eta_j} \left[\int_0^{\min\{T-t_j;\varepsilon\}} \frac{\partial}{\partial \boldsymbol{\alpha}} g_{\boldsymbol{\alpha}}(t|\kappa_j) \, \mathrm{d}t \right] \left[\int_{R_j} f_{\boldsymbol{\sigma}}(\boldsymbol{s}|\kappa_j) \, \mathrm{d}\boldsymbol{s} \right] .$$

Web Appendix B: Fisher Information Matrix

The inverse of the Fisher information matrix at the maximum likelihood estimate (MLE) $\hat{\boldsymbol{\theta}}_{ML}$ is in general likelihood theory used as an estimate of the variance of $\hat{\boldsymbol{\theta}}_{ML}$. The precise conditions under which asymptotic properties of MLEs hold for spatio-temporal point processes have been established by Rathbun (1996). Specifically, the conditions for existence, consistence and asymptotic normality of a local maximum $\hat{\boldsymbol{\theta}}_{ML}$ as $T \to \infty$ for a fixed observation region W are discussed in Meyer (2009, Section 4.2.3).

The expected Fisher information $\mathcal{I}(\boldsymbol{\theta})$ can be estimated by the "optional variation process" adapted to the marked spatio-temporal setting (Rathbun, 1996, equation (4.7))

$$\int_0^T \int_W \int_{\mathcal{K}} \left(\frac{\partial}{\partial \boldsymbol{\theta}} \log \lambda^*_{\boldsymbol{\theta}}(t, \boldsymbol{s}, \kappa) \right)^{\otimes 2} \mathrm{d}N(t, \boldsymbol{s}, \kappa)$$

through its observed realisation

$$\hat{\mathcal{I}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \left(\frac{\partial}{\partial \tilde{\boldsymbol{\theta}}} \log \lambda^*_{\tilde{\boldsymbol{\theta}}}(t_i, \boldsymbol{s}_i, \kappa_i) \Big|_{\tilde{\boldsymbol{\theta}}=\boldsymbol{\theta}} \right)^{\otimes 2} = \sum_{i=1}^{n} \left(\frac{\frac{\partial}{\partial \tilde{\boldsymbol{\theta}}} \lambda^*_{\tilde{\boldsymbol{\theta}}}(t_i, \boldsymbol{s}_i, \kappa_i)}{\lambda^*_{\tilde{\boldsymbol{\theta}}}(t_i, \boldsymbol{s}_i, \kappa_i)} \Big|_{\tilde{\boldsymbol{\theta}}=\boldsymbol{\theta}} \right)^{\otimes 2},$$

where $\boldsymbol{a}^{\otimes 2} := \boldsymbol{a}\boldsymbol{a}'$ for a vector $\boldsymbol{a}$. Uncertainty of the parameter estimates is thus deduced from the diagonal of $\hat{\mathcal{I}}^{-1/2}(\hat{\boldsymbol{\theta}}_{ML})$, which contains their standard errors.

References

Meyer, S. (2009). Spatio-temporal infectious disease epidemiology based on point processes. Master's thesis, Department of Statistics, Ludwig-Maximilians-Universität, München. Available as `http://epub.ub.uni-muenchen.de/11703/`.

Rathbun, S. L. (1996). Asymptotic properties of the maximum likelihood estimator for spatio-temporal point processes. *Journal of Statistical Planning and Inference* **51,** 55–74.

Web Appendix C: Simulation Algorithm

This appendix provides a more implementational view on the simulation algorithm described in Section 4 of the paper. In addition to the notation of the paper, let $L(t)$ be the next time point after time t where any endemic covariate in any tile changes its value, or a previously infected individual stops spreading the disease, i.e.:

$$L(t) = \min\left\{t' > t \mid (\exists\, \xi \in \{1,\dots,M\} : \boldsymbol{z}_{\tau(t'),\xi} \neq \boldsymbol{z}_{\tau(t),\xi}) \vee (\exists\, j \in \{1,\dots,N_g(t)\} : t' = t_j + \varepsilon)\right\}.$$

An implementational perspective on the simulation algorithm is then:

Algorithm 1: Ogata's modified thinning adapted for `twinstim`

1. Given current time t, update $L(t)$ and calculate local upper bound $\overline{\lambda_g^*}(t)$;
2. Generate proposed waiting time $\Delta \sim \mathrm{Exp}(\overline{\lambda_g^*}(t))$;
3. **if** $t + \Delta > L(t)$ **then**
4. Let $t = L(t)$;
5. **else**
6. Let $t = t + \Delta$;
7. Accept t with probability $\lambda_g^*(t)/\overline{\lambda_g^*}(t)$, otherwise goto step 1;
8. Draw the source of infection (infective individual j or endemic) with weights equal to the respective components of $\lambda_g^*(t)$ (cf. equation (8) of the paper);
9. **if** *endemic source of infection* **then**
10. Draw the type κ of the new event with weights $\exp(\beta_0(\kappa))$, $\kappa \in \mathcal{K}$;
11. Draw the tile A_ξ of the new event with weights $|A_\xi|\, \rho_{\tau(t),\xi}\, e^{\boldsymbol{\beta}' \boldsymbol{z}_{\tau(t),\xi}}$, $\xi \in \{1,\dots,M\}$;
12. Draw the location $\boldsymbol{s}$ of the new event uniform within the sampled tile A_ξ;
13. **else**
14. Draw the type κ of the new event at random out of the types which can be triggered by the source individual j, i.e. draw from $U(\{\kappa \in \mathcal{K} : q_{\kappa_j,\kappa} = 1\})$;
15. Draw the relative location $\boldsymbol{v}$ of the new event (relative to the source j) from the density $f(\boldsymbol{s}|\kappa_j)/\int_{R_j} f(\boldsymbol{s}|\kappa_j)\,\mathrm{d}s$ on R_j, i.e. $\boldsymbol{s} = \boldsymbol{s}_j + \boldsymbol{v}$;
16. Draw additional marks according to the pre-specified distribution;
17. Update the event history;
18. Goto step 1;

An `R` implementation of the algorithm can be found as the function `simEpidataCS` in the popular `surveillance` package.

APPENDIX B

Flexible estimation of spatio-temporal interaction in a point process model for infectious disease spread

Sebastian Meyer, Leonhard Held

Extended abstract published in the *Proceedings of the 29th International Workshop on Statistical Modelling*, Göttingen, Germany, 2014.

Flexible estimation of spatio-temporal interaction in a point process model for infectious disease spread

Sebastian Meyer[1], Leonhard Held[1]

[1] University of Zurich, Switzerland

E-mail for correspondence: Sebastian.Meyer@ifspm.uzh.ch

Abstract: Case reports from infectious disease surveillance with registered location and time of infection allow for spatio-temporal point process models of infectious disease spread. An endemic component describes the baseline risk of infection driven by population density as well as temporal and exogenous effects. A second, epidemic component captures interaction between cases and includes covariate effects on the force of infection. Here we investigate nonparametric estimation of spatial as well as temporal interaction using B-splines. Such flexible formulations disclose the distance-decay and time-course of infectivity in a more data-driven manner than previously used parametric models.

Keywords: infectious disease surveillance; self-exciting spatio-temporal point process with immigration; conditional intensity model; interaction function.

1 Introduction

Infectious disease surveillance aims at the timely detection of outbreaks as well as their prevention and control. Public health authorities routinely collect data on the occurrence of communicable diseases, registering location and date of infection, and the specific pathogen involved. The case reports often contain patient characteristics which are potentially associated with individual infectivity, e.g. the patient's age, and are ideally supplemented by lattice data on environmental and socio-demographic factors for spatial regression. Given such surveillance data, spatio-temporal point process models are a useful tool to estimate the role of individual characteristics and exogenous factors in shaping disease spread.

Following Held et al. (2005), disease risk is decomposed additively into two components: An *endemic* component describes the baseline risk driven by population density as well as temporal and exogenous effects (e.g., seasonality, population structure, prevalence of correlated diseases), whereas an

This paper was published as a part of the proceedings of the 29th International Workshop on Statistical Modelling, Georg-August-Universität Göttingen, 14–18 July 2014.

observation-driven *epidemic* component invokes explicit dependence between cases. Meyer et al. (2012) proposed a spatio-temporal point process model with such components, and applied it to 635 cases of invasive meningococcal disease (IMD) caused by the two most common meningococcal finetypes in Germany,2002–2008. They identified a time trend with seasonal pattern, and no evidence for an additional (lagged) association with local waves of influenza. The epidemic component revealed that the meningococcus of serogroup B was approximately twice as infectious as the C-type. However, spatial and temporal interaction of cases were modelled rather naively by assuming a *Gaussian* distance decay of *time-constant* infectivity. The former was improved by Meyer and Held (2014), who found power laws for spatial interaction to outperform previous formulations in two modelling frameworks for infectious disease surveillance data. In this paper, we investigate the use of B-splines to estimate interaction in a more flexible way.

2 Self-exciting spatio-temporal point process model

The spatio-temporal point process model proposed by Meyer et al. (2012) is designed for time-space-mark data $\{(t_i, \mathbf{s}_i, \mathbf{m}_i) : i = 1, \ldots, n\}$ of dependent events such as case reports of infectious diseases. It is defined through the conditional intensity function

$$\lambda(t, \mathbf{s}) = \nu_{[t][\mathbf{s}]}\, \rho_{[t][\mathbf{s}]} + \sum_{j:t_j<t} \eta_j \cdot g(t - t_j) \cdot f(\|\mathbf{s} - \mathbf{s}_j\|) \tag{1}$$

in a region $\mathbf{W} \ni \mathbf{s}$ during a period $(0, T] \ni t$. The first, endemic component consists of a log-linear predictor $\log(\nu_{[t][\mathbf{s}]}) = \beta_0 + \beta^{\mathrm{T}}\mathbf{z}_{[t][\mathbf{s}]}$ with a multiplicative offset $\rho_{[t][\mathbf{s}]}$, typically the population density. Both the offset and exogenous covariates in $\nu_{[t][\mathbf{s}]}$ are given on a spatio-temporal grid, hence the notation $[t][\mathbf{s}]$ for the period containing t in the region covering $\mathbf{s}$. Note that such a piecewise constant endemic model is equivalent to a Poisson regression model for the aggregated number of cases on the given grid.
However, with the second, epidemic component the intensity process depends on previously infected individuals and becomes "self-exciting". The epidemic force of infection at $(t, \mathbf{s})$ is the superposition of the infection pressures caused by previously infected individuals. The log-linear predictor $\log(\eta_j) = \gamma_0 + \gamma^{\mathrm{T}}\mathbf{m}_j$ weights infectivity by individual/infection-specific characteristics $\mathbf{m}_j$. Regional-level covariates from the endemic grid can also be included in η_j, e.g., to model ecological effects on infectivity.
Decreasing infection pressure over space is described by $f(x)$ as a function of the spatial distance from the infectious source. Meyer et al. (2012) originally used the Gaussian kernel $f_{\mathrm{G}}(x) = \exp\left\{-x^2/(2\sigma^2)\right\}$ as a standard choice. Subsequently, Meyer and Held (2014) showed that the power law

$$f_{\mathrm{PL}}(x) = (x + \sigma)^{-d}\,, \tag{2}$$

$\sigma, d > 0$, is more appropriate in describing distance decay of infectivity, which seems to translate from the power-law feature of human travel (Brockmann et al., 2006). For the time course of infectivity, $g(t)$, both works simply assumed a constant function over a fixed period.
However, the basic model framework (1) actually allows for arbitrary shapes with the only requirements that the interaction functions are differentiable with respect to their parameters, and that $g(t)$ and $f_2(\mathbf{s}) = f(\|\mathbf{s}\|)$ are integrable over $(0, T]$ and $\mathbf{W} - \mathbf{s}_j$, respectively. In what follows, we investigate flexible estimates of spatial and temporal interaction for the IMD data, retaining the endemic component and η_j from the previous analyses.

3 Spatial interaction

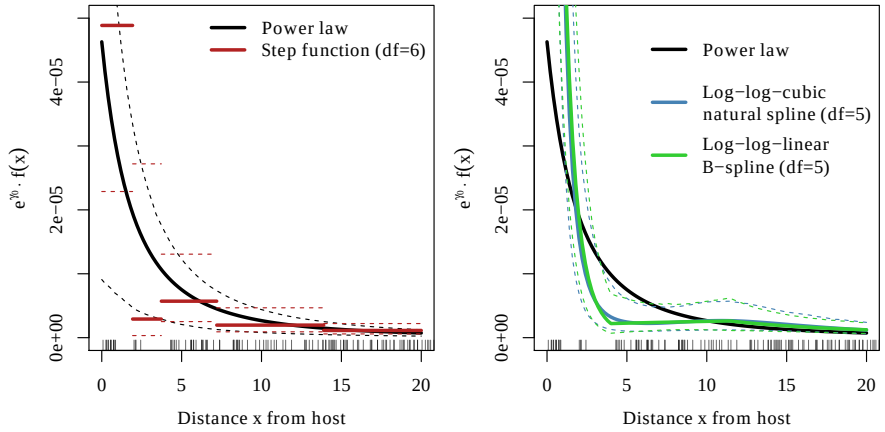


FIGURE 1. Flexible estimates of spatial interaction vs. the power law. Dashed lines represent 95% confidence intervals and the bottom "rug" shows the observed distances of events to their potential sources, i.e., to events of the past 30 days.

The left part of Figure 1 shows results from Meyer and Held (2014) with fixed $g(t) = \mathbb{I}_{(0,30]}(t)$. The estimated power law features a pronounced initial decay of infectivity as well as a heavy tail capturing occasional transmissions over large distances. The step function was estimated with six log-equidistant knots up to an upper-bound range of 100 kilometres. It suggests an even sharper initial drop but forfeits monotonicity.
A more sophisticated approach of flexible estimation is a log-log B-spline

$$f_{\mathrm{B}}(x) = \exp\left\{\sum_{k=1}^{K} \alpha_k B_k(\log(x+\sigma))\right\}, \tag{3}$$

where the B_k form a set of suitable basis functions (Fahrmeir et al., 2013, Section 8.1). The log-log formulation is motivated by the fact that the power law (2) turns into a simple linear relation on that scale:

$$\log(e^{\gamma_0} \cdot f_{\mathrm{PL}}(x)) = \gamma_0 - d \cdot \log(x+\sigma)\,.$$

This is also why linear basis functions might be sufficiently flexible, although resulting in non-differentiable joints. The right plot of Figure 1 shows estimates based on a linear and a natural cubic B-spline, respectively, each with 5 degrees of freedom. We fixed σ at the estimate from the power-law model ($\sigma = 4.60$), and the inner knots were again chosen to be equidistant on the log-scale. The spline fits are very similar and in accordance with the step function in suggesting an even steeper initial decay than the power law. With respect to AIC they perform substantially better: $\Delta\text{AIC} = -23.7$ for the cubic, and -22.2 for the linear variant, respectively. Note, however, that computational cost of the B-spline models is more than 10-fold compared to the power law. We have to evaluate basis functions and, most notably, we cannot simplify the spatial integrals $\int_{\mathbf{W}-\mathbf{s}_j} f(\|\mathbf{s}\|)\,\mathrm{d}\mathbf{s}$ in the likelihood to a one-dimensional quadrature problem (cf. Meyer and Held, 2014, Supplement B), but have to rely on product Gauss cubature.

4 Temporal interaction

Temporal interaction $g(t)$ has not been estimated in previous models for the IMD data, but assumed constant for 30 days from infection, vanishing to zero afterwards. This reluctance is mainly due to the sparseness of cases: on average, there are only 4 and 3.6 infections with types B and C, respectively, per month. For illustration, we estimate some alternative temporal interaction functions $g(t)$ while sticking to the upperbound length of 30 days for the infectious period and employing the spatial power law $f_{\text{PL}}(x)$.

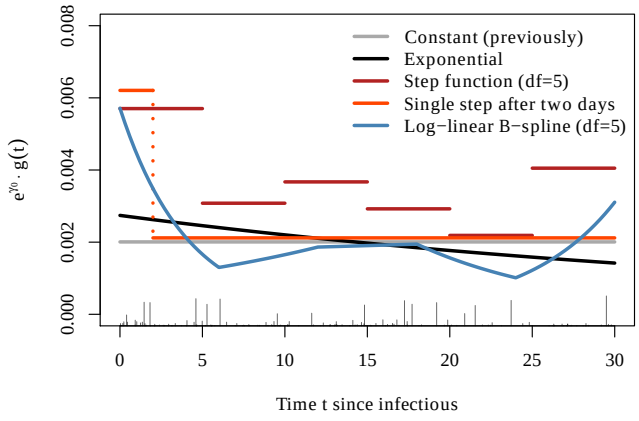


FIGURE 2. Point estimates of various models for temporal interaction. The bottom "rug" shows the observed time lags between events, where the size corresponds to the associated spatial interaction given by the estimated power law from Figure 1.

A simple parametric model for the time course of infectivity is exponential decay $g_{\text{E}}(t) = e^{-\alpha t}$, $\alpha > 0$. The B-spline formulation (3) is also applicable

for $g(t)$ but using plain rather than log-scale. Figure 2 shows estimates of temporal interaction assuming either exponential decay, or B-splines of degrees 0 (step function) or 1 with equidistant knots, or a simplified step function with a single knot after 2 days. Note that the overall level varies slightly between the alternatives since a change in $g(t)$ also affects the estimation of $f_{\mathrm{PL}}(x)$ and γ_0. One would expect the more flexible estimates to approach zero for larger time lags. However, considering late infections close to previously infective sites tends to improve the likelihood, since the endemic component is only constant within districts. It is thus necessary to determine a reasonable range of temporal interaction by other means, e.g., epidemiological considerations.
Concerning model performance, only the simplified one-step function improves upon the previous constant model ($\Delta\mathrm{AIC} = -6.4$). It suggests a partial drop of infectivity already after two days, which might correspond to quarantine actions taken after the appearance of symptoms.

5 Conclusion

We have shown that flexible B-spline formulations of interaction can be incorporated into a spatio-temporal point process model for infectious disease surveillance data. They may deliver additional insight into the spatial dependence structure and the time course of infectivity. It is crucial that the spatio-temporal resolution of the surveillance data is high enough to allow for flexible estimation of interaction. The IMD data set used for illustration is rather small and carries little information at small distances, which is why results should be regarded with caution.
As a common drawback, the regression splines depend on the chosen (boundary) knots and require much more computation time, especially in the spatial domain. A compromise are 0-degree B-splines, which don't require numerical integration and may serve as a quick initial benchmark for spatial and temporal interaction.
The application of this and two related model frameworks in R is described in detail in Meyer et al. (2014).
Software: All calculations have been carried out in the statistical software environment R 3.0.3. The point process model (1) is implemented in the R package **surveillance** as function `twinstim()`, and a simplified version of the IMD data is included as `data("imdepi")` (courtesy of the German Reference Centre for Meningococci). Spatial integrals in the likelihood have been evaluated using cubature methods from the R package **polyCub** (see Meyer and Held, 2014, Supplement B). The implementation also allows for other specifications of the interaction functions f and g, respectively.

Acknowledgments: The research is financially supported by the Swiss National Science Foundation (project 137919: *Statistical methods for spatio-temporal modelling and prediction of infectious diseases*).

References

Brockmann, D., Hufnagel, L., and Geisel, T. (2006). The scaling laws of human travel. *Nature*, **439**, 462 – 465.

Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. (2013). Regression: Models, Methods and Applications. Berlin: Springer-Verlag.

Held, L., Höhle, M., and Hofmann, M. (2005). A statistical framework for the analysis of multivariate infectious disease surveillance counts. *Statistical Modelling*, **5**, 187 – 199.

Meyer, S., Elias, J., and Höhle, M. (2012). A space-time conditional intensity model for invasive meningococcal disease occurrence. *Biometrics*, **68**, 607 – 616.

Meyer, S. and Held, L. (2014). Power-law models for infectious disease spread. *Annals of Applied Statistics*. Accepted and available as `arXiv:1308.5115`.

Meyer, S., Held, L., and Höhle, M. (2014). Spatio-temporal modeling of epidemic phenomena using the R package **surveillance**. In preparation for the *Journal of Statistical Software*.